Reproducibility and dependencies for Jupyter Notebooks

Operate First Data Science Community Meetup, 30th November 2021

Francesco Murdaca, Red Hat

Red Hat

Agenda

Jupyter Notebooks (~2 min)

A quick intro to Jupyter Notebooks.

What problem are we trying to solve? (~6 min)

Dependency management for Jupyter Notebooks.

Project Thoth (~4 min)

Project Thoth overview

How does Thoth help to solve the problem? (~5 min)

How Thoth contributes to the solution of the problems stated.

Dependency Management Tutorial (~5 min)

Operate First, Project Meteor and dependency management tutorial

Conclusion (~3 min)

Jupyter Notebook

Operate First Data Science Community Meetup

Red Hat

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

Language of choice

Jupyter support over 40 programming languages.

Notebook sharing

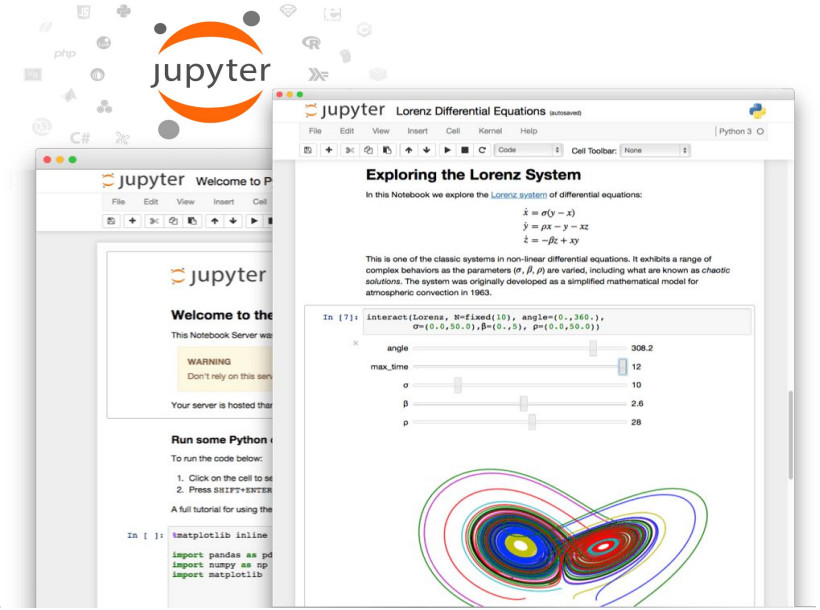
Sharing interactive code documents with others.

Interactive output

Rich, interactive output: HTML, images, videos, etc.

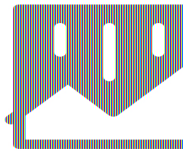
Big Data integration and analysis

Leverage big data tools and explore that data.



Source https://jupyter.org/

Heavily adopted



Data Scientists

Data analysis, modeling and visualization and analytical reports.

Educators and students

Assignments, interactive coding lessons, tutorials.

Developers

Rapid prototyping, POCs, testing and integration and example usage.

Trusted by **many**

Red Hat

What problem are we trying to solve?

Operate First Data Science Community Meetup

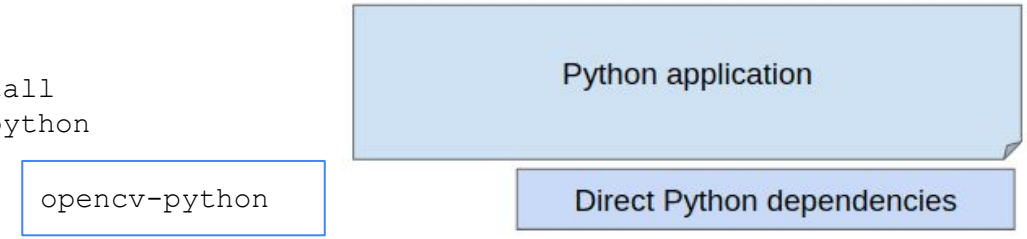
Red Hat

What problem are we trying to solve?

```
pip install
opencv-python
```

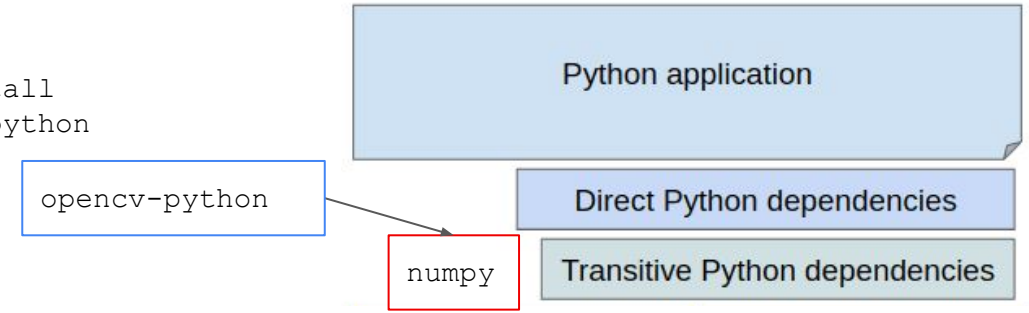
What problem are we trying to solve?

```
pip install
opencv-python
```



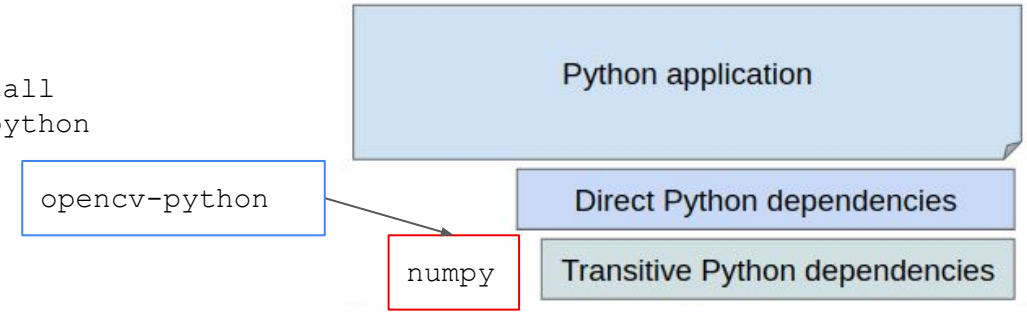
What problem are we trying to solve?

```
pip install
opencv-python
```



What problem are we trying to solve?

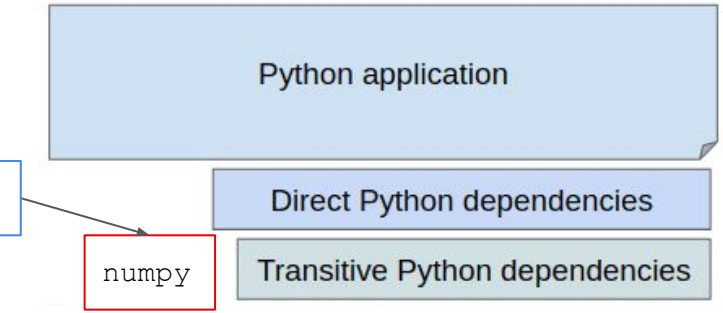
```
pip install
opencv-python
```



What about versions?

What problem are we trying to solve?

```
pip install
opencv-python
```



Releases 57

3.4.16.57 Latest

5 days ago

Releases 72

v1.21.3 Latest

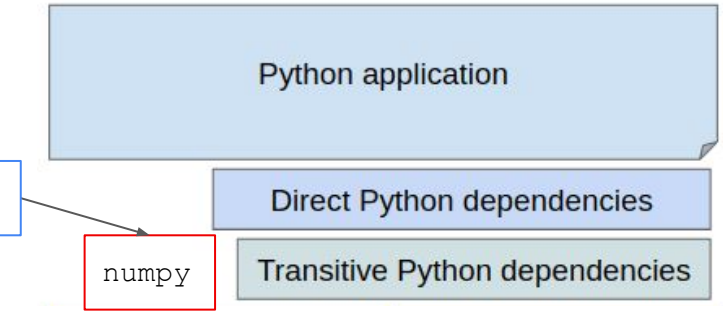
6 days ago

What about versions?

Red Hat

What problem are we trying to solve?

```
pip install
opencv-python
```

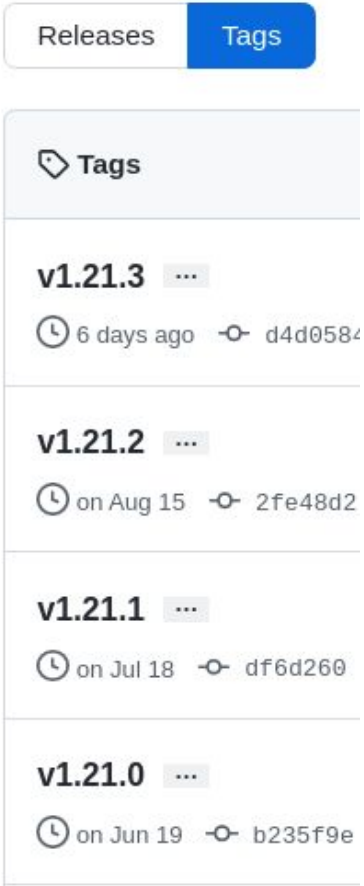


Releases 57

3.4.16.57 Latest
5 days ago

Releases 72

v1.21.3 Latest
6 days ago

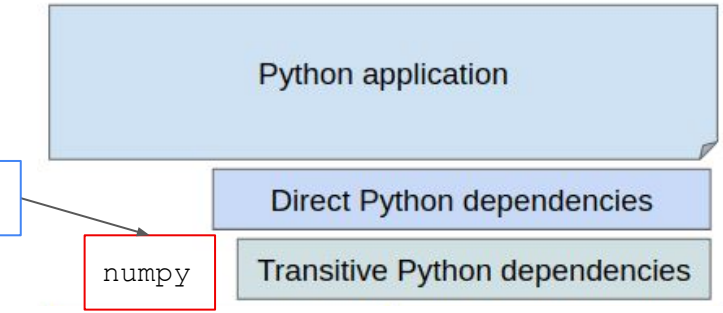


What about versions?

Red Hat

What problem are we trying to solve?

```
pip install
opencv-python
```

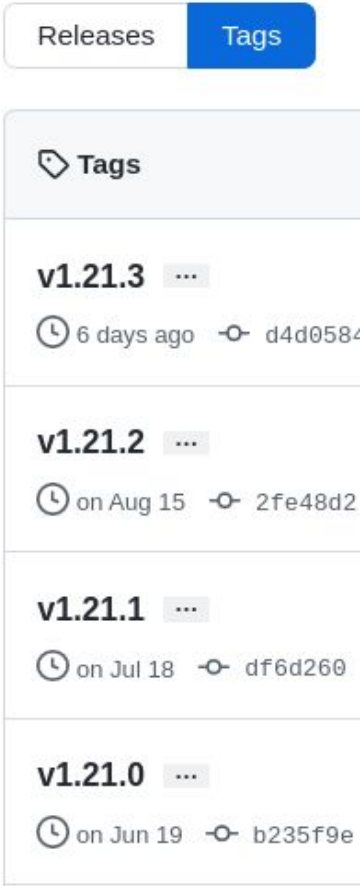


Releases 57

3.4.16.57 Latest
5 days ago

Releases 72

v1.21.3 Latest
6 days ago

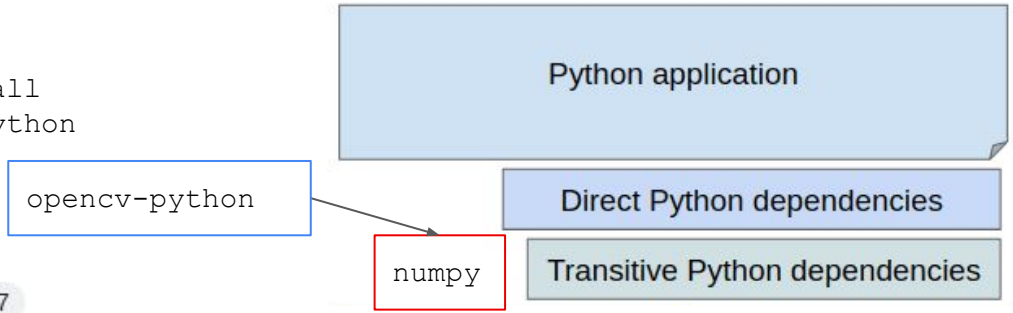


What about versions?
What about hashes?

Red Hat

What problem are we trying to solve?

```
pip install
opencv-python
```

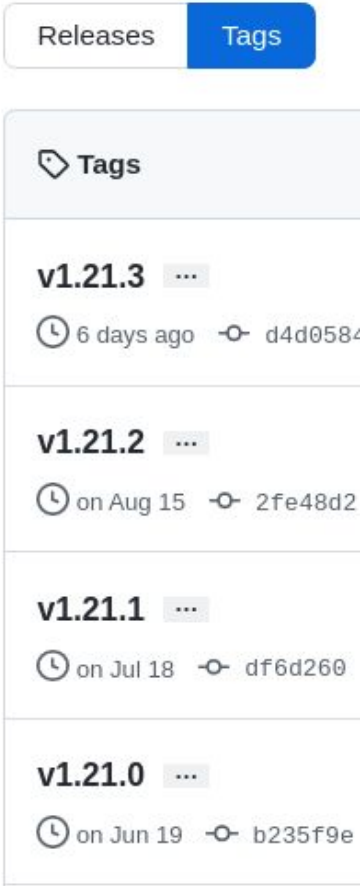


Releases 57

3.4.16.57 Latest
5 days ago

Releases 72

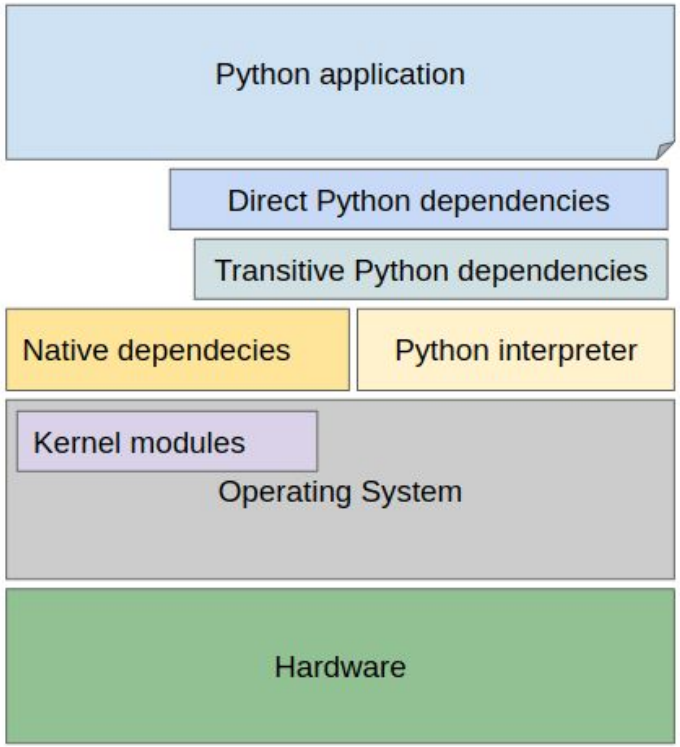
v1.21.3 Latest
6 days ago



What about versions?
What about hashes?
What about Python interpreter?

Red Hat

What problem are we trying to solve?



What problem are we trying to solve?

Install dependencies

In [2]:
```
! pip install tensorflow
! pip install boto3
! pip install matplotlib
```

What problem are we trying to solve?

Install dependencies

```
In [2]: ! pip install tensorflow
        ! pip install boto3
        ! pip install matplotlib
```

This will not guarantee reproducibility!

What problem are we trying to solve?

```
voila
folium
numpy
pandas
ipywidgets
ipykernel
matplotlib
```

What problem are we trying to solve?

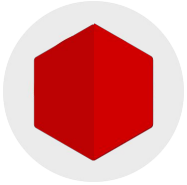
```
voila
folium
numpy
pandas
ipywidgets
ipykernel
matplotlib
```

Having requirements.txt with no versions stated does not guarantee to have reproducible notebook!

Jupyter Notebooks are by default **NOT stand-alone**

Managing dependencies

Requirements
are **decoupled from a notebook*** into
manifest files, such as
requirements.txt or *Pipfile.lock*



Containerisation

A specialised tools or a custom Dockerfile is needed so that all notebooks requirements* are present in the resulting image.

Sharing

The consumer must first **manually** set up **environment** for them using provided* **manifest** files.

*It is not uncommon that **NO manifest files are provided** and hence Notebook users must **find out dependencies themselves**.

Source: https://www.linkedin.com/pulse/jupyter-notebooks-production-marek-%C4%8Derm%C3%A1k/

Difficulties for both authors and consumers

Authors have to...

Create an environment

Ideally a virtual environment for the notebook to run in.

Install dependencies to the environment

[optional] Create/Update custom kernel

It is a recommended approach (and the best practice) to create a custom kernel for each project.

[optional] Create/Update manifest files

Consumers have to...

Create an environment

Ideally a virtual environment for the notebook to run in.

Install dependencies* to the environment

[optional] Create custom kernel

It is a recommended approach (and the best practice) to create a custom kernel for each project.

*It is not uncommon that **NO manifest files are provided** and hence Notebook users must **find out dependencies themselves**.

Source: https://www.linkedin.com/pulse/jupyter-notebooks-production-marek-%C4%8Derm%C3%A1k/

Difficulties for both authors and consumers

Authors have to...

Create an environment

Ideally a virtual environment for the notebook to run in.

Install dependencies to the environment

[optional] Create/Update custom kernel

It is a recommended approach (and the best practice) to create a custom kernel for each project.

[optional] Create/Update manifest files

Consumers have to...

Create an environment

Ideally a virtual environment for the notebook to run in.

Install dependencies* to the environment

[optional] Create custom kernel

It is a recommended approach (and the best practice) to create a custom kernel for each project.

*It is not uncommon that **NO manifest files are provided** and hence Notebook users must **find out dependencies themselves**.

Source: https://www.linkedin.com/pulse/jupyter-notebooks-production-marek-%C4%8Derm%C3%A1k/

Difficulties for both authors and consumers

Authors have to...

Create an environment

Ideally a virtual environment for the notebook to run in.

Install dependencies to the environment

[optional] Create/Update custom kernel

It is a recommended approach (and the best practice)
to create a custom kernel for each project.

[optional] Create/Update manifest files

Consumers have to...

Create an environment

Ideally a virtual environment for the notebook to run in.

Install dependencies* to the environment

[optional] Create custom kernel

It is a recommended approach (and the best practice)
to create a custom kernel for each project.

Reproducibility!

*It is not uncommon that **NO manifest files are provided** and hence Notebook users must **find out dependencies themselves**.

Source: https://www.linkedin.com/pulse/jupyter-notebooks-production-marek-%C4%8Derm%C3%A1k/

Project Thoth

Operate First Data Science Community Meetup

Red Hat

Project Thoth

- Provide **a system and a user facing service**, that helps making well educated decisions by delivering a broad and deep knowledge set wrt frameworks relevant in the field of **AI applications**.
- **Deliver optimized, secured, well maintained and predictable images** for your AI applications
- Use **bots to automate mundane work** to offload humans work

Source: https://thoth-station.ninja/

Thoth Recommendation types

- Latest
- Stable
- Security
- Performance
- Testing

Source: https://thoth-station.ninja/

what we observe and store in our knowledge graph

Application Stack related:

- Buildtime and runtime environment
- Dependencies
- Performances

Software Package:

- Application Binary Interfaces (ABI)
- Security (CVE, analyzers)

Source Code Meta Information:

- Project features (TTR, TTCI, etc,..) from different software development platform (Github, GitLab, Pagure, etc...)

Source: https://thoth-station.ninja/

what we observe and use in resolution process

Prescriptions:

- Pipeline units described using a YAML file to heal Python Application
  - Pipeline unit type and its configuration used in the resolution process
- Opens a possibility to easily create pipeline units specific for recurrent issues or use-cases following a pattern
  - e.g. a pipeline unit specific for ODH use case when resolving for ODH base container images
  - ...

Check the video

Source: https://github.com/thoth-station/prescriptions

How do we use this knowledge?

- Recommender system is called **Adviser** in Thoth.
- It uses Reinforcement Learning (RL).

Check the video

Source: https://thoth-station.ninja/

Thoth Integrations

- Command line tool thamos (developer laptop)
- Cyborg Kebechet (pull request/issues creator)
- Source-to-Image (container builder)
- Optimizing Deployment Pipelines
- **Jupyter Tools (data scientist browser)**

Source: https://thoth-station.ninja/

How does Thoth help solve the problem?

Operate First Data Science Community Meetup

Red Hat

jupyterlab-requirements

JupyterLab extension for dependency management and optimization

JupyterLab extension allows you to **manage dependencies and store everything in the Jupyter notebook metadata**:

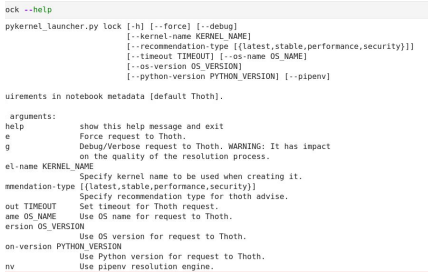
- **requirements (Pipfile)**;
- **requirements locked** with all versions and hashes of libraries (direct and transitive ones) **(Pipfile.lock)**;
- **dependency resolution engine** used **(Thoth or Pipenv)**;
- configuration file containing runtime environment (only for Thoth resolution engine).

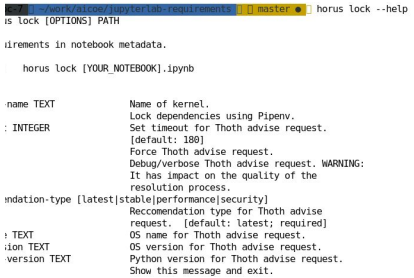
Source:https://pypi.org/project/jupyterlab-requirements/

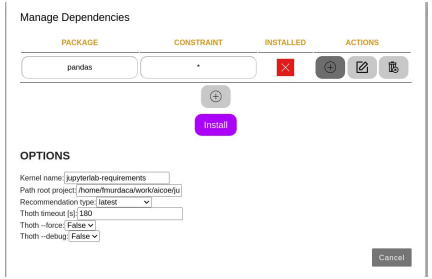
How does Thoth helps solving the problem?

jupyterlab-requirements

JupyterLab extension for dependency management and optimization







Jupyter magic commands

available directly in your notebook cells when you start a notebook

CLI

that you can run from terminal or integrated in pipelines

UI

accessible through **Manage Dependencies button** that appears in the notebook when it is opened in JupyteLab

Source:https://pypi.org/project/jupyterlab-requirements/

Red Hat

jupyterlab-requirements

Jupyter magic commands

```
[2]: %horus lock --help

usage: ipykernel_launcher.py lock [-h] [--force] [--debug]
                                  [--kernel-name KERNEL_NAME]
                                  [--recommendation-type [{latest,stable,performance,security}]]
                                  [--timeout TIMEOUT] [--os-name OS_NAME]
                                  [--os-version OS_VERSION]
                                  [--python-version PYTHON_VERSION] [--pipenv]

Lock requirements in notebook metadata [default Thoth].

optional arguments:
  -h, --help            show this help message and exit
  --force               Force request to Thoth.
  --debug               Debug/Verbose request to Thoth. WARNING: It has impact
                        on the quality of the resolution process.
  --kernel-name KERNEL_NAME
                        Specify kernel name to be used when creating it.
  --recommendation-type [{latest,stable,performance,security}]
                        Specify recommendation type for thoth advise.
  --timeout TIMEOUT     Set timeout for Thoth request.
  --os-name OS_NAME     Use OS name for request to Thoth.
  --os-version OS_VERSION
                        Use OS version for request to Thoth.
  --python-version PYTHON_VERSION
                        Use Python version for request to Thoth.
  --pipenv              Use pipenv resolution engine.
```

- Run commands from notebook cells:
  - Handle dependencies from cells (add/remove)
  - Handle kernels from cells (set-kernel)
  - Lock dependencies from cells (lock)
- Different resolution engines (Thoth, Pipenv)

Source:https://pypi.org/project/jupyterlab-requirements/

jupyterlab-requirements

Command Line Interface (CLI)

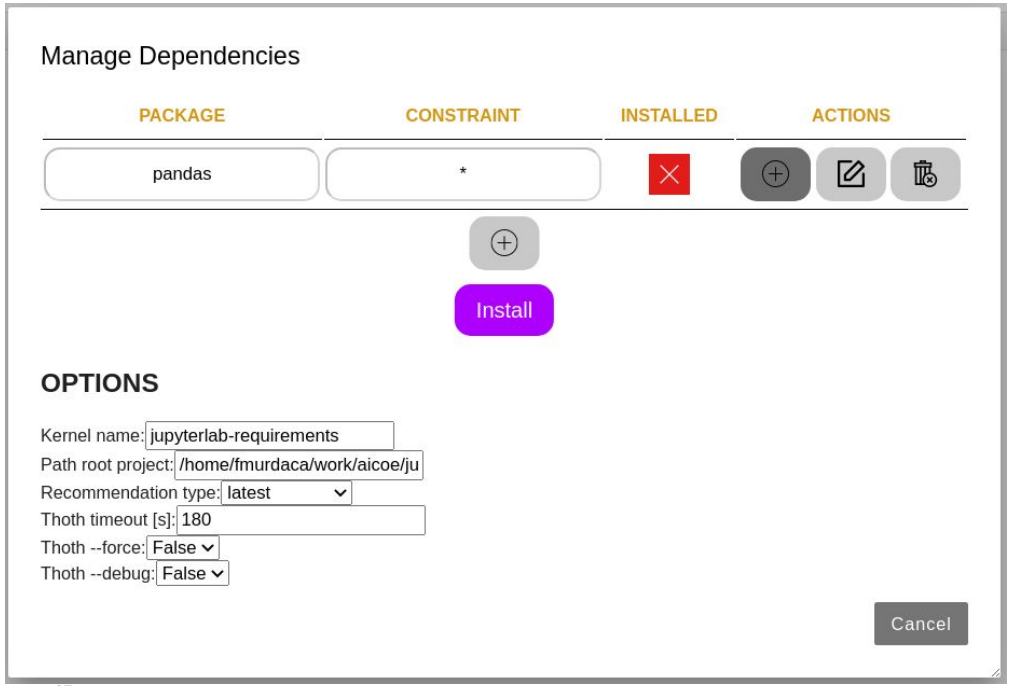
```
fmurdaca@pc-7  ~/work/aicoe/jupyterlab-requirements  master ●  horus lock --help
Usage: horus lock [OPTIONS] PATH

  Lock requirements in notebook metadata.

  Examples:   horus lock [YOUR_NOTEBOOK].ipynb

Options:
  --kernel-name TEXT              Name of kernel.
  --pipenv                        Lock dependencies using Pipenv.
  --timeout INTEGER               Set timeout for Thoth advise request.
                                  [default: 180]
  --force                         Force Thoth advise request.
  --debug                         Debug/verbose Thoth advise request. WARNING:
                                  It has impact on the quality of the
                                  resolution process.
  --recommendation-type [latest|stable|performance|security]
                                  Reccomendation type for Thoth advise
                                  request.  [default: latest; required]
  --os-name TEXT                  OS name for Thoth advise request.
  --os-version TEXT               OS version for Thoth advise request.
  --python-version TEXT           Python version for Thoth advise request.
  --help                          Show this message and exit.
```

- Run commands from terminal
  - Handle dependencies from cells (add/remove)
  - Handle kernels from cells (set-kernel)
  - Lock dependencies from cells (lock)
- Handle jupyter notebook dependencies in CI/CD pipelines
- Different resolution engines (Thoth, Pipenv)

Source:https://pypi.org/project/jupyterlab-requirements/

jupyterlab-requirements

User Interface (UI)



- Interactive UI to handle dependencies

Source: https://pypi.org/project/jupyterlab-requirements/

jupyterlab-requirements Python package

```
pip install jupyterlab-requirements

jupyter lab
```


Source: https://pypi.org/project/jupyterlab-requirements/

Dependency management tutorial for Jupyter Notebooks

Operate First Data Science Community Meetup

Red Hat

Dependency management tutorial

- start working on a new notebook
- create dependencies for your existing notebook
- convert notebook that uses pip commands in cells
- use a reproducible notebook

Source: https://github.com/AICoE/manage-dependencies-tutorial

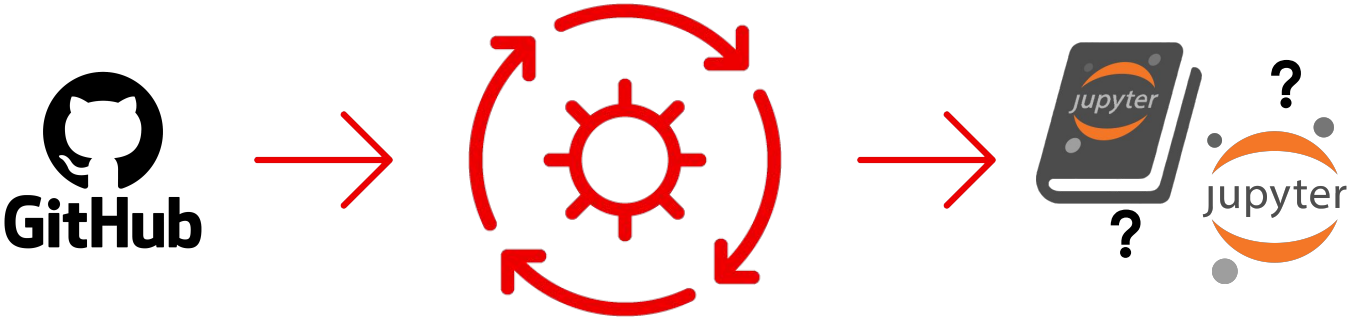
Operate First

https://operate-first.cloud

https://github.com/operate-first

Source: https://www.operate-first.cloud/

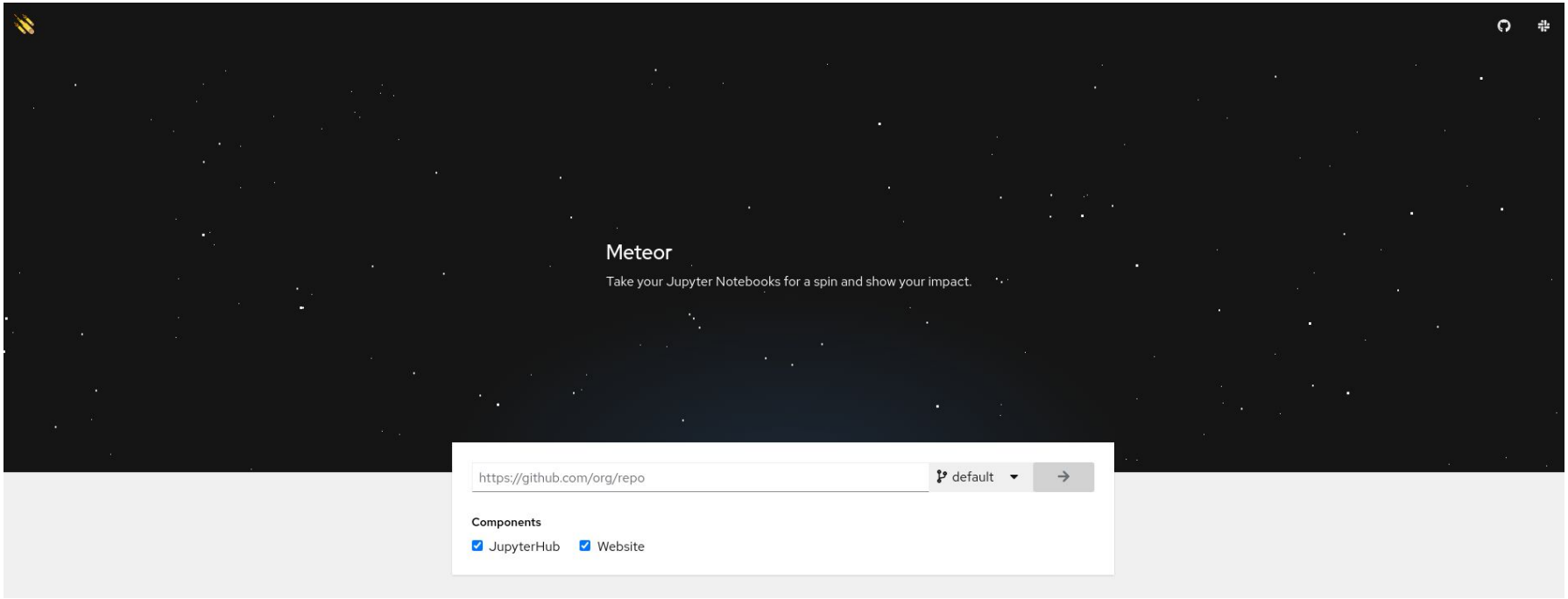
Project Meteor



AICoE CI,
Tekton pipelines,
Thoth Advise

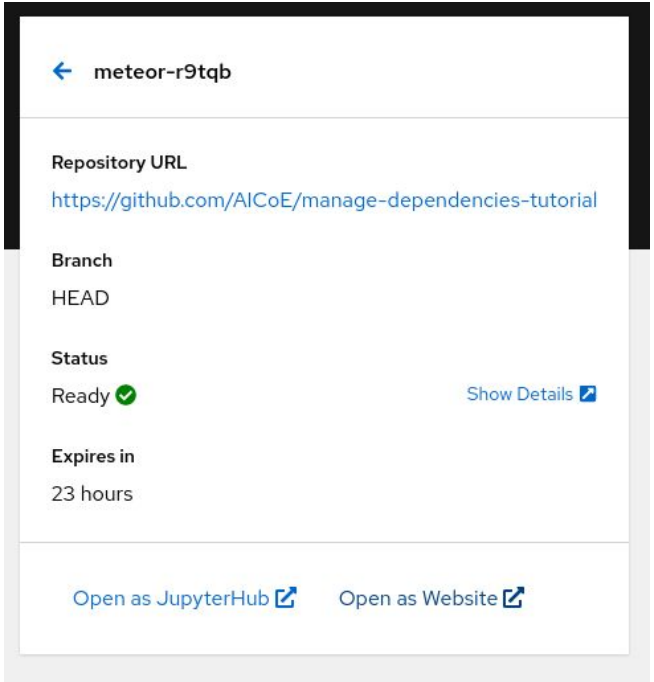
Source: https://github.com/AICoE/meteor

Project Meteor



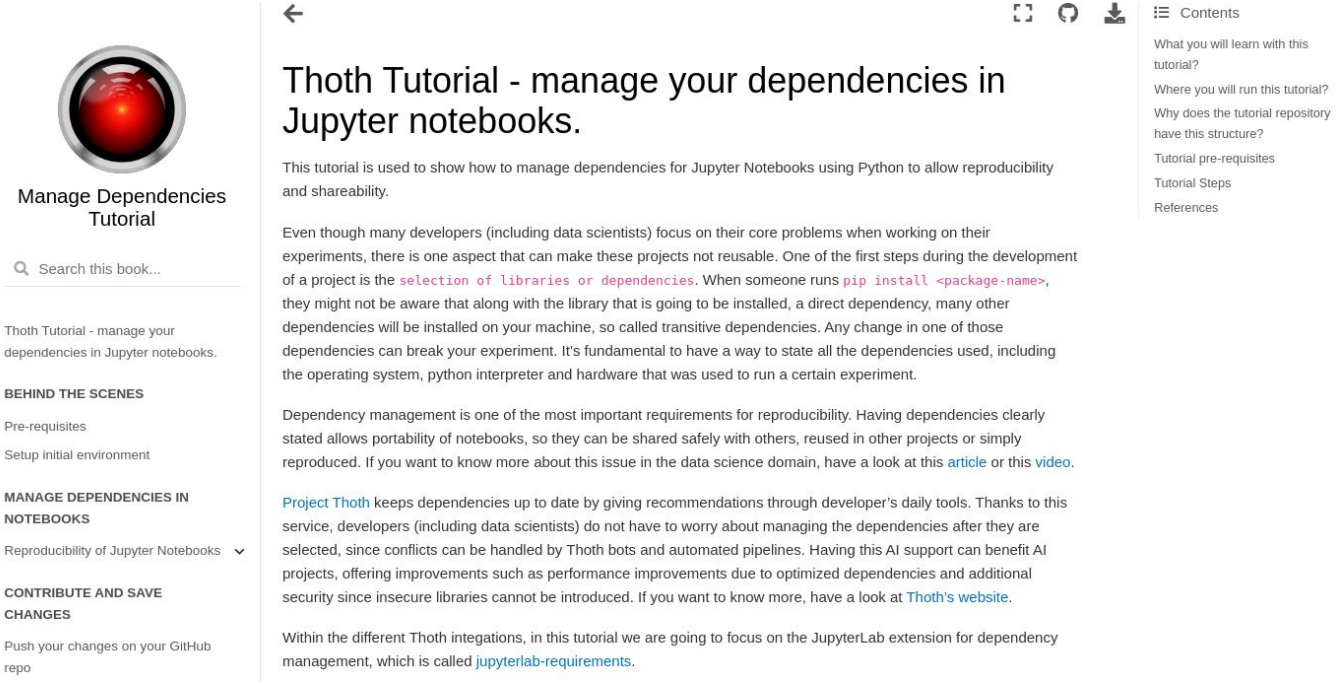
Source: https://shower.meteor.zone/

Project Meteor



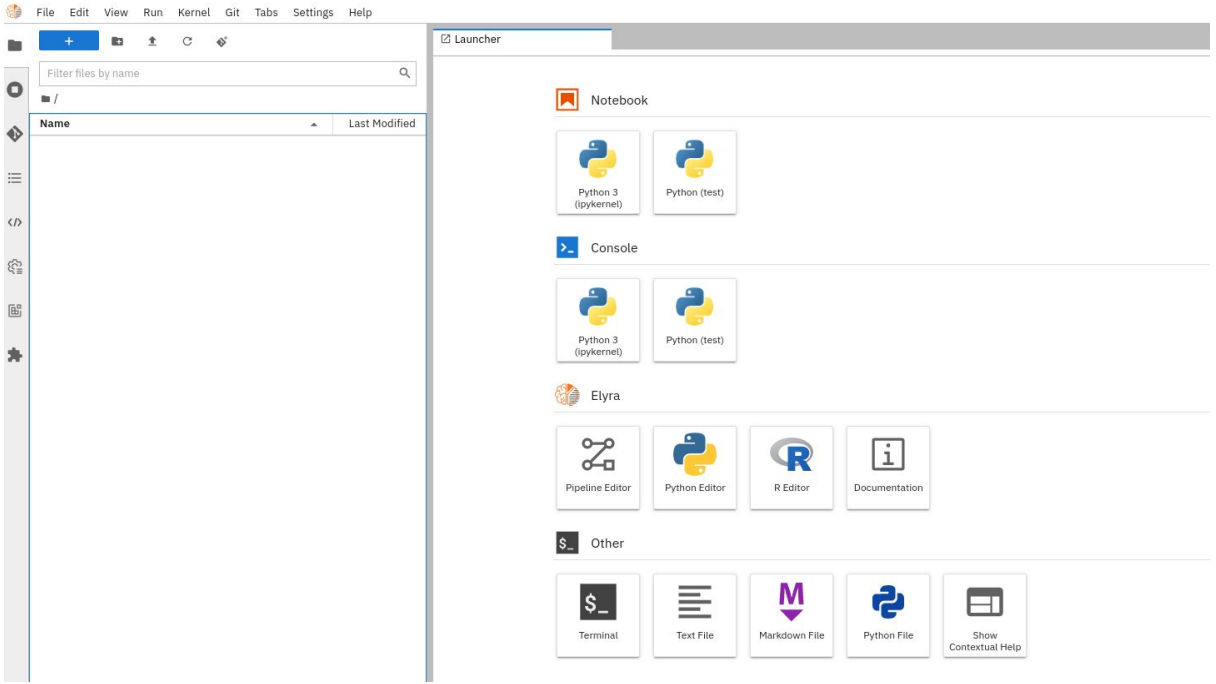
Source: https://github.com/AICoE/meteor

Project Meteor - Jupyter Book



Source: https://github.com/AICoE/meteor

Project Meteor - JupyterLab environment



Source: https://github.com/AICoE/meteor

Conclusions

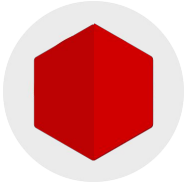
Operate First Data Science Community Meetup

A notable improvements...

Managing dependencies

~~Requirements are **decoupled from a notebook*** into manifest files, such as *requirements.txt* or *Pipfile.lock*~~

Requirements are **locked** and **embedded** directly into the notebook. No additional files are needed.



Containerisation

~~A specialised tools or a custom Dockerfile is needed so that all notebooks requirements* are present in the resulting image.~~

Jupyter Notebooks with embedded dependencies can be built directly using **Jupyter Notebook S2I without any additional files**.

Sharing

~~The consumer must first **manually** set up **environment** for them using provided* **manifest** files.~~

Jupyter Notebooks can be shared as **stand-alone units** without any additional files. Environment is prepared in a **single click**.

Red Hat

With the focus on reproducibility

Resolved Jupyter Notebook dependencies

When the notebook is distributed, unless specified otherwise, the **very same versions** are used which guarantees compatibility and reliable results.*

*Arguably, reproducibility is not only about same versions, other factors have to be considered as well, like hardware and architecture, or random variables.

Project Thoth

- **Website** https://thoth-station.ninja/
- **Twitter** https://twitter.com/thothstation
- **GitHub** https://github.com/thoth-station
- **Thoth Station YouTube** https://www.youtube.com/channel/UCIUIDuq_hQ6vlzmqM59B2Lw/videos
- **Tutorial** https://github.com/AICoE/manage-dependencies-tutorial

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat