openclean

Open-Source Data Profiling and Data Cleaning Library

https://github.com/VIDA-NYU/openclean

Motivation

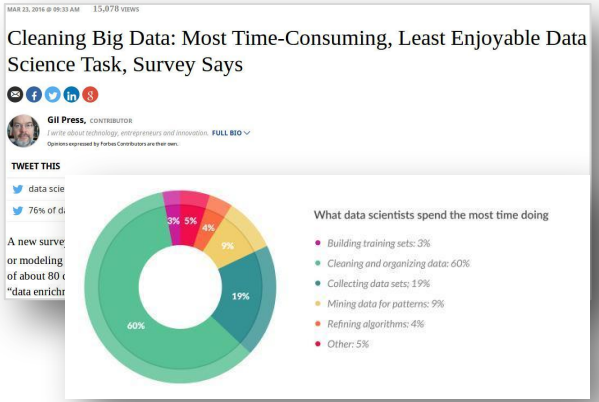
openclean

Data cleaning is a major bottleneck for many data science projects.

Many tools for profiling and cleaning data have been developed, both in academia and industry.

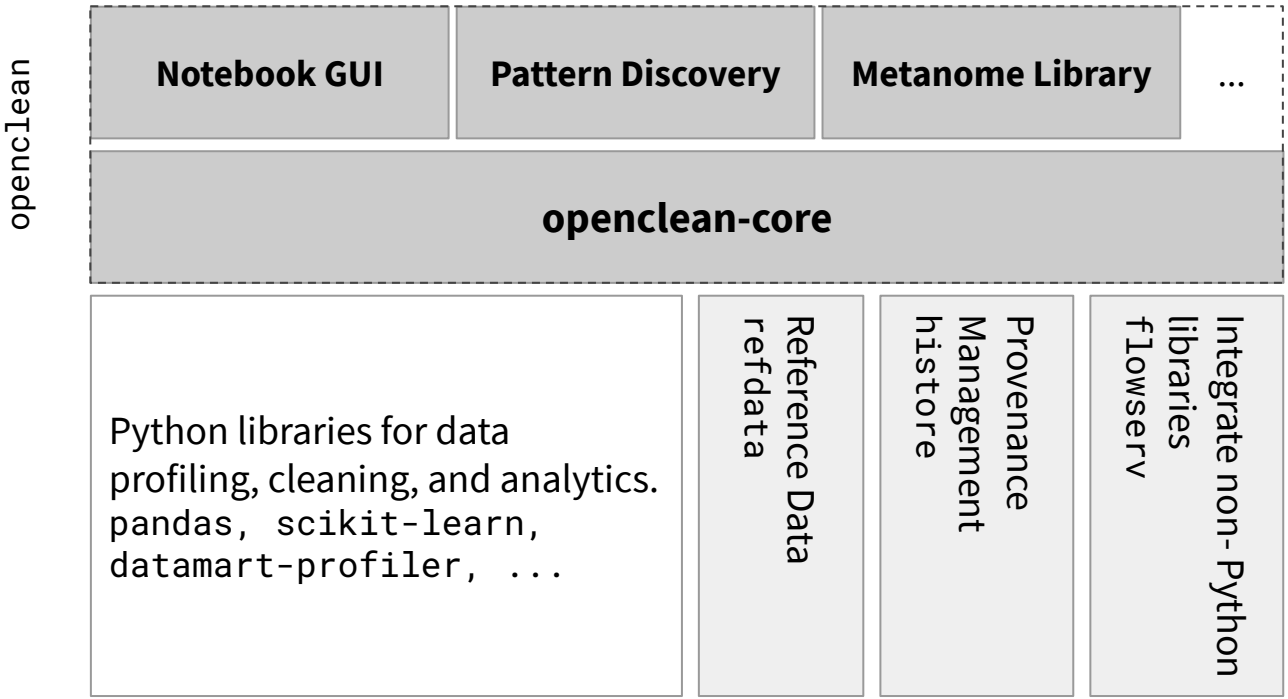
openclean is a open-source Python library that …

- integrates existing tools in single environment,
- is easy to use and extensible, and
- builds a community of users, developers, and researchers.



data curation

openclean



data curation

openclean

Install and Run the Demo Notebooks

1

```
# Create a new virtual environment
virtualenv venv
source venv/bin/activate

-- or --

# Virtual environment using conda
conda create -n openclean pip
conda activate openclean
```

2

```
# Clone openclean repository
git clone git@github.com:VIDA-NYU/openclean.git

# Change working directory
cd openclean

# Install openclean and dependencies
pip install .[demo]

# Run Jupyter (navigate to examples/notebooks)
jupyter notebook
```

data curation