renku Intro

Rok Roškar, Swiss Data Science Center, ETH Zürich

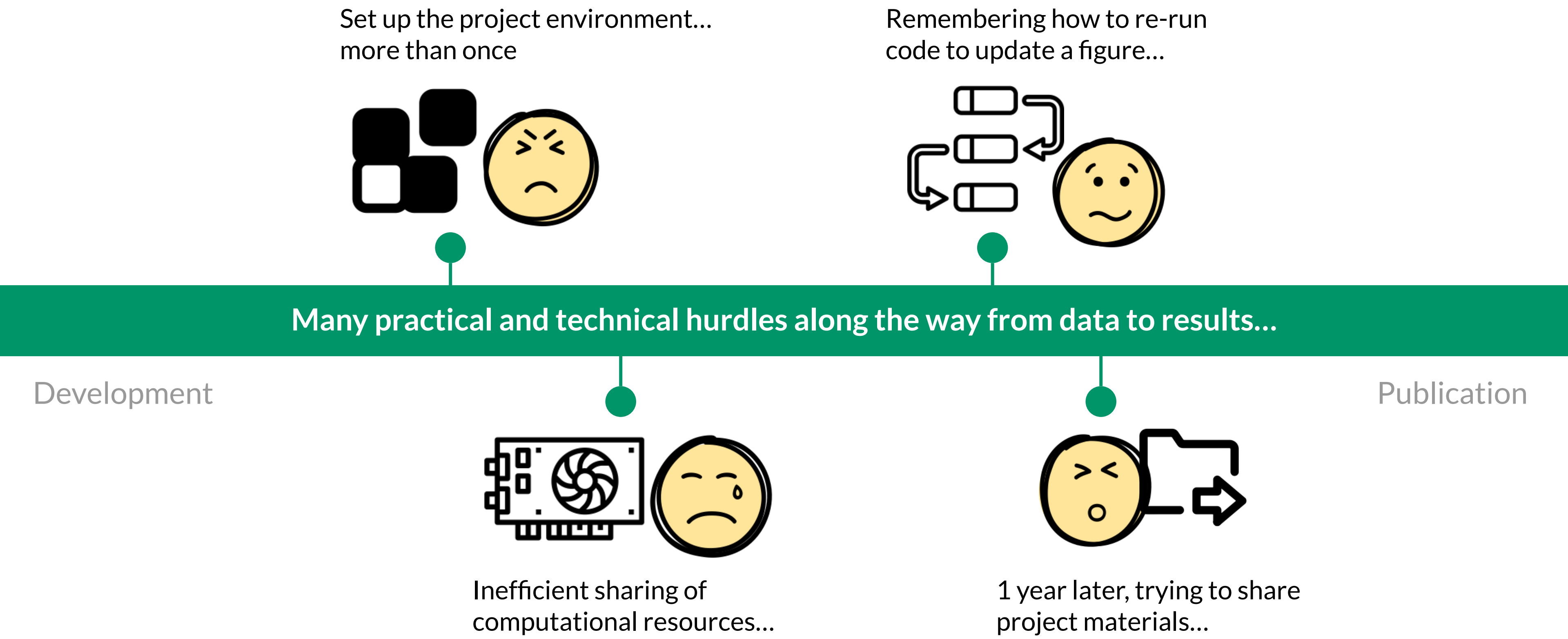
September 6, 2023

SDSC EPFL ETH zürich

Renku enables “sustainable” data science

1. Done **today** with **tomorrow** in mind
2. Individual work benefits **institution** and **community**
3. All components form a functioning **ecosystem**

a story…



Set up the project environment… more than once

Remembering how to re-run code to update a figure…

Many practical and technical hurdles along the way from data to results…

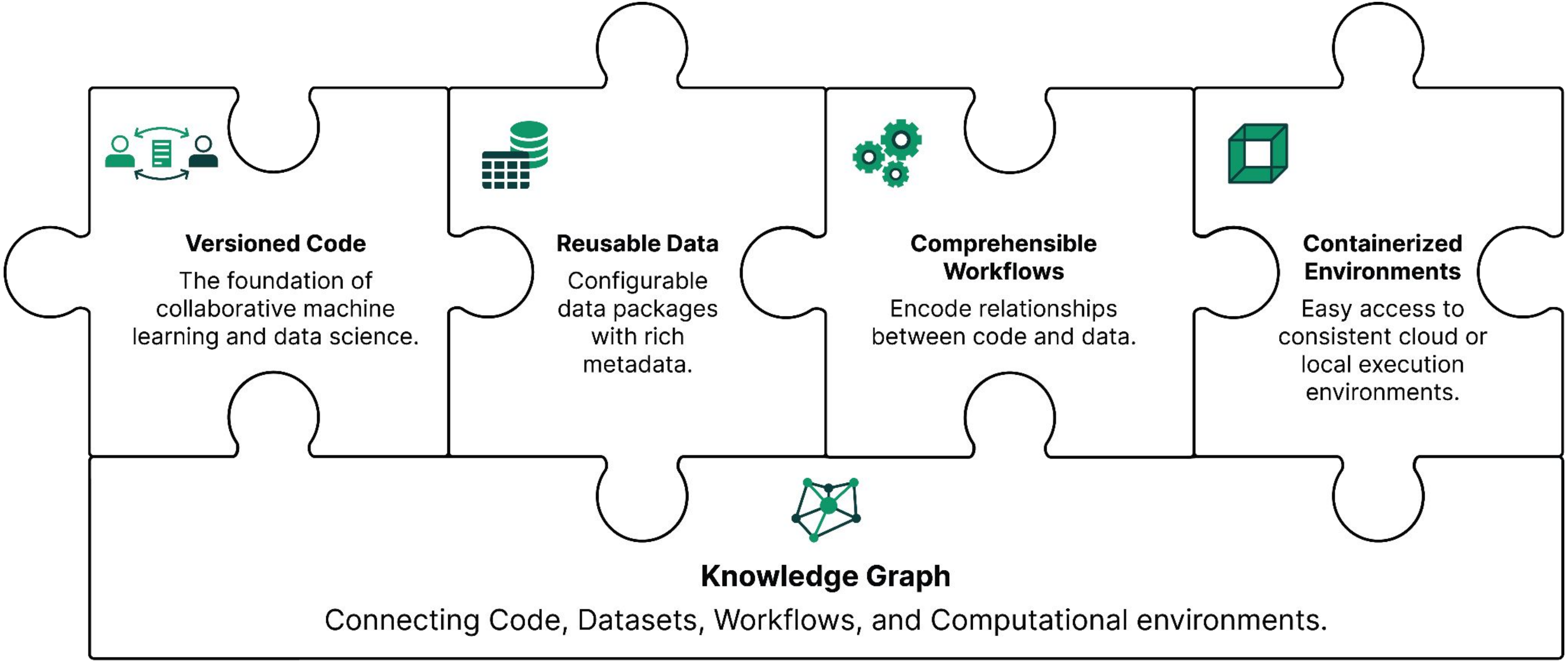
Development

Publication

Inefficient sharing of computational resources…

1 year later, trying to share project materials…

What is Renku?



Versioned Code
The foundation of collaborative machine learning and data science.

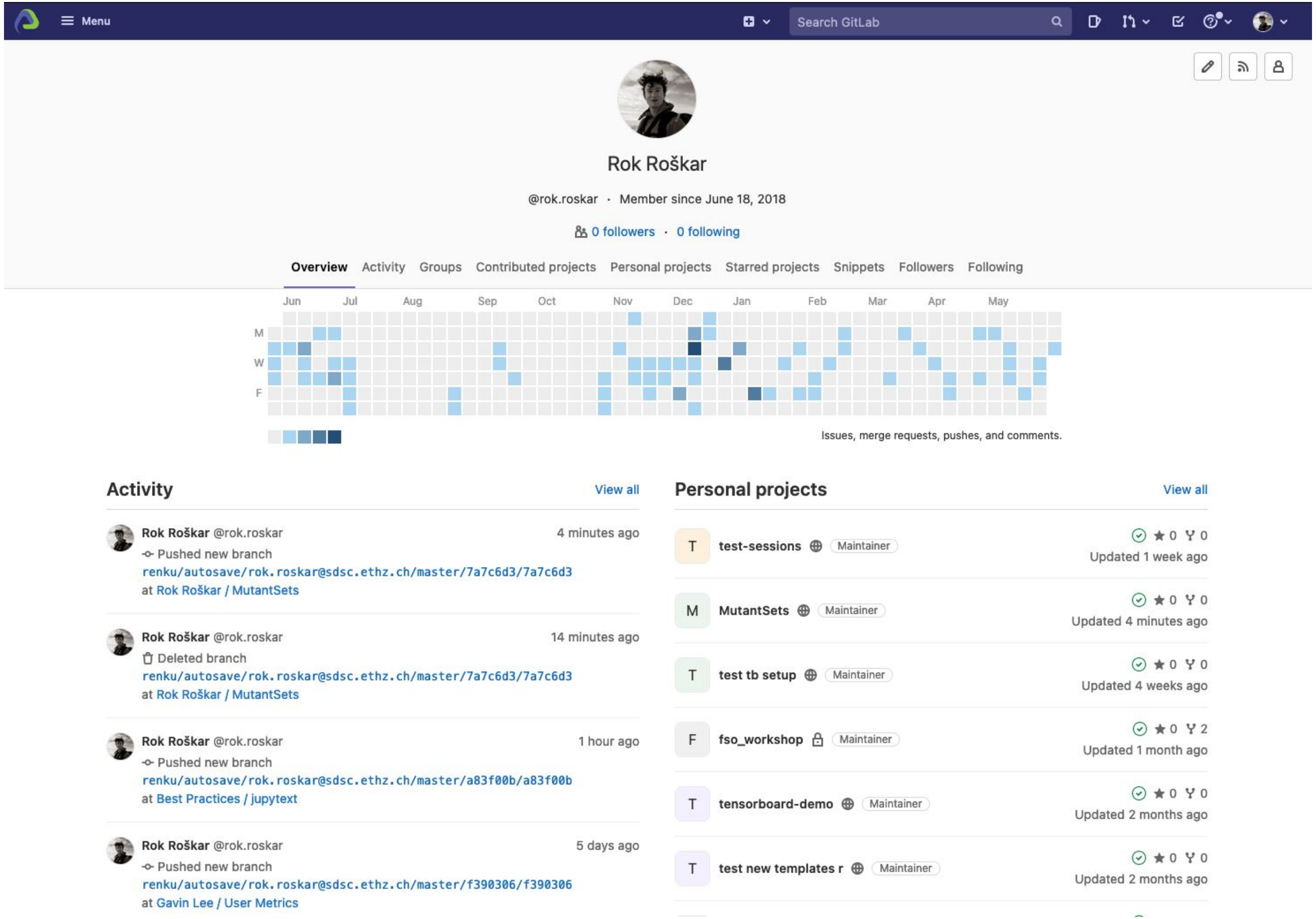
Reusable Data
Configurable data packages with rich metadata.

Comprehensible Workflows
Encode relationships between code and data.

Containerized Environments
Easy access to consistent cloud or local execution environments.

Knowledge Graph
Connecting Code, Datasets, Workflows, and Computational environments.

Code



Rok Roškar

@rok.roskar · Member since June 18, 2018

0 followers · 0 following

Issues, merge requests, pushes, and comments.

Activity

View all

Rok Roškar @rok.roskar 4 minutes ago
Pushed new branch
renku/autosave/rok.roskar@sdsc.ethz.ch/master/7a7c6d3/7a7c6d3
at Rok Roškar / MutantSets

Rok Roškar @rok.roskar 14 minutes ago
Deleted branch
renku/autosave/rok.roskar@sdsc.ethz.ch/master/7a7c6d3/7a7c6d3
at Rok Roškar / MutantSets

Rok Roškar @rok.roskar 1 hour ago
Pushed new branch
renku/autosave/rok.roskar@sdsc.ethz.ch/master/a83f00b/a83f00b
at Best Practices / jupytext

Rok Roškar @rok.roskar 5 days ago
Pushed new branch
renku/autosave/rok.roskar@sdsc.ethz.ch/master/f390306/f390306
at Gavin Lee / User Metrics

Personal projects

View all

T test-sessions Maintainer ★ 0 ⑂ 0 Updated 1 week ago

M MutantSets Maintainer ★ 0 ⑂ 0 Updated 4 minutes ago

T test tb setup Maintainer ★ 0 ⑂ 0 Updated 4 weeks ago

F fso_workshop Maintainer ★ 0 ⑂ 2 Updated 1 month ago

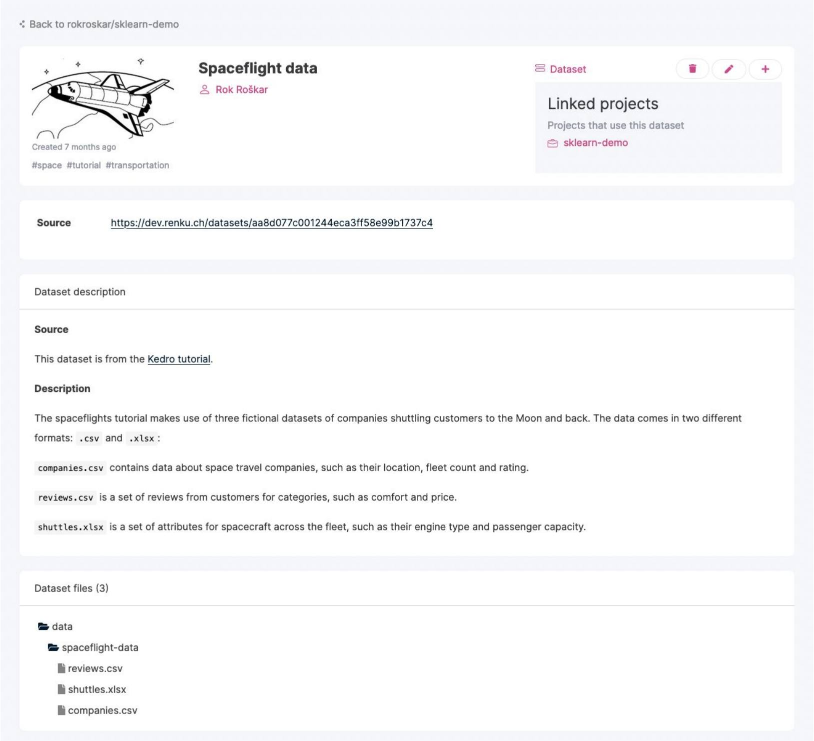
T tensorboard-demo Maintainer ★ 0 ⑂ 0 Updated 2 months ago

T test new templates r Maintainer ★ 0 ⑂ 0 Updated 2 months ago

Datasets

Create (choose storage), assemble, annotate, publish

- Create datasets to easily reuse and share data across projects
- Use various backends: git-LFS, S3, Azure blob, local or network storage sources
- Combine with compute environments and analysis examples to ensure data can easily be used and reused
- Record pipelines that yield or consume datasets for full traceability



Back to rokroskar/sklearn-demo

Spaceflight data

Rok Roškar

Created 7 months ago

#space #tutorial #transportation

Dataset

Linked projects

Projects that use this dataset

sklearn-demo

Source https://dev.renku.ch/datasets/aa8d077c001244eca3ff58e99b1737c4

Dataset description

Source

This dataset is from the Kedro tutorial.

Description

The spaceflights tutorial makes use of three fictional datasets of companies shuttling customers to the Moon and back. The data comes in two different formats: `.csv` and `.xlsx`:

`companies.csv` contains data about space travel companies, such as their location, fleet count and rating.

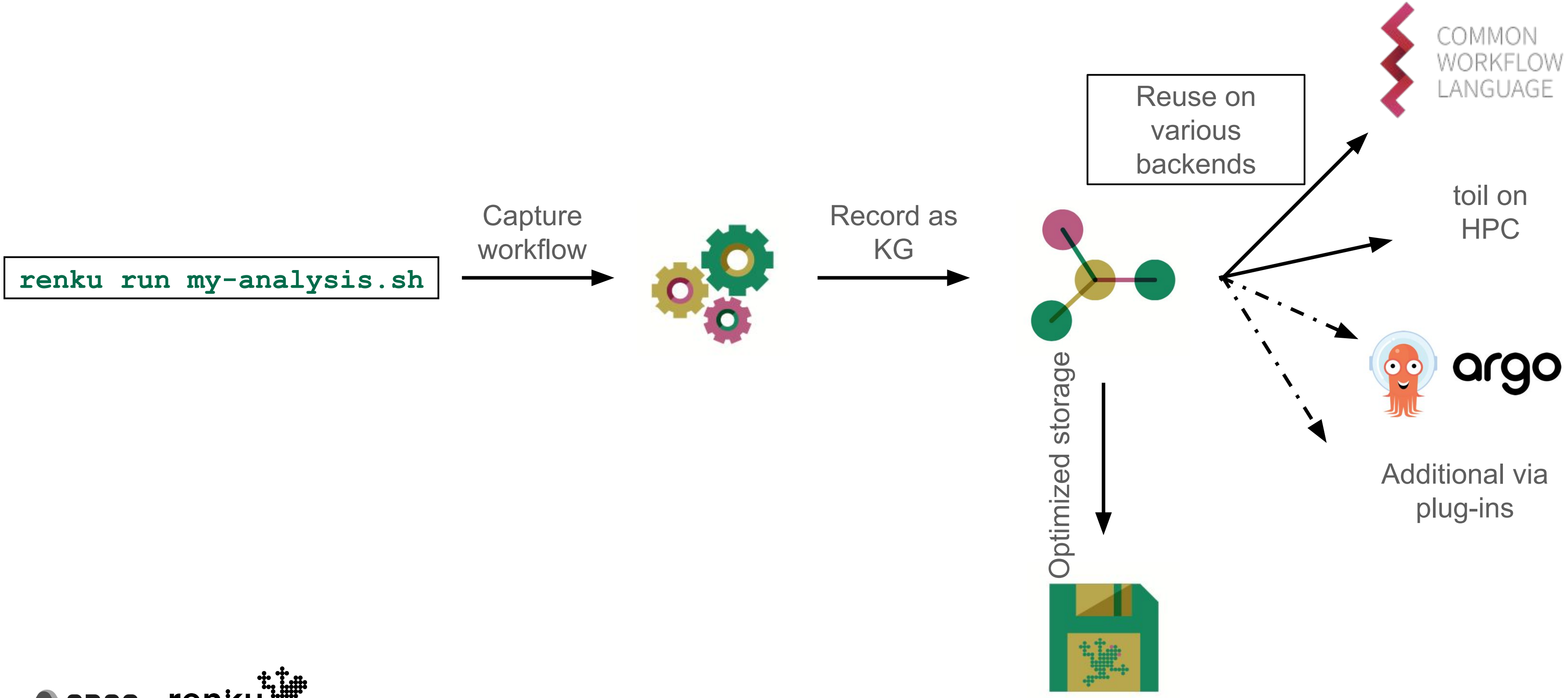
`reviews.csv` is a set of reviews from customers for categories, such as comfort and price.

`shuttles.xlsx` is a set of attributes for spacecraft across the fleet, such as their engine type and passenger capacity.

Dataset files (3)

```
data
  spaceflight-data
    reviews.csv
    shuttles.xlsx
    companies.csv
```

Workflows



```
renku run my-analysis.sh
```

Capture workflow

Record as KG

Optimized storage

Reuse on various backends

toil on HPC

Additional via plug-ins

Automatically record pipelines

Legend

Code files

code_file.py

Code in a programming language like python or R

Script execution

python filename.py

The command which is run, typically taking input arguments and producing output results.

Data files

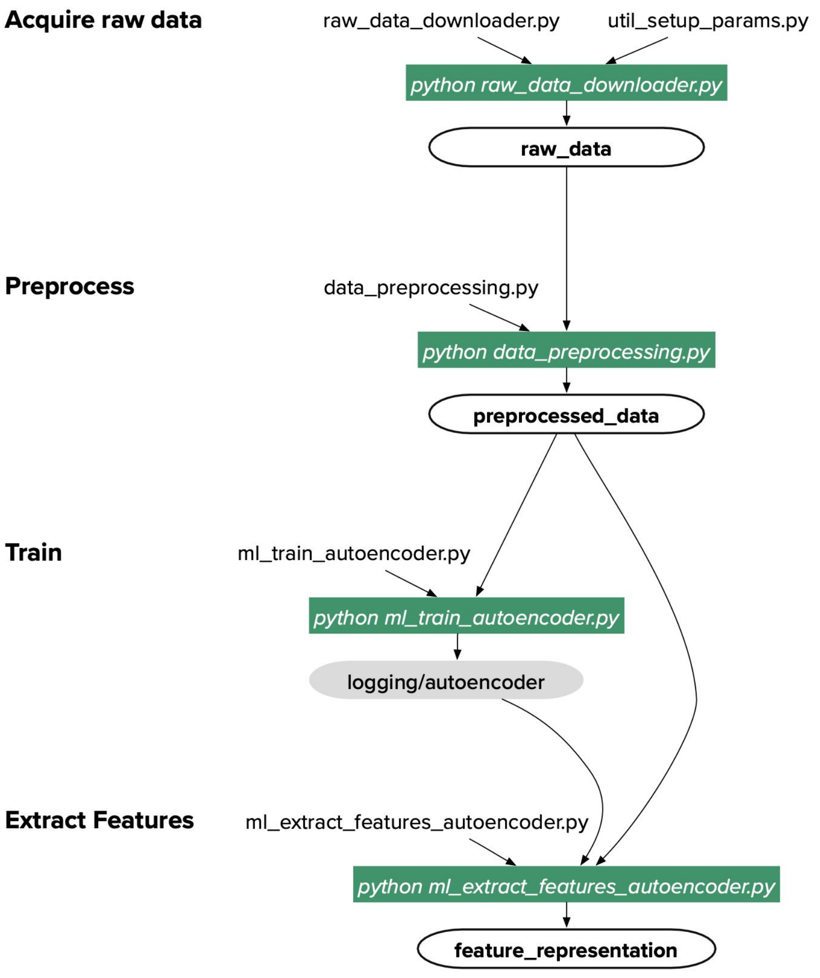
data_file.ext

A file or folder containing data.

Datasets

dataset name

A collection of files and/or folders that contain data and metadata describing information like authorship, licensing, etc.



Acquire raw data

raw_data_downloader.py

util_setup_params.py

python raw_data_downloader.py

raw_data

Preprocess

data_preprocessing.py

python data_preprocessing.py

preprocessed_data

Train

ml_train_autoencoder.py

python ml_train_autoencoder.py

logging/autoencoder

Extract Features

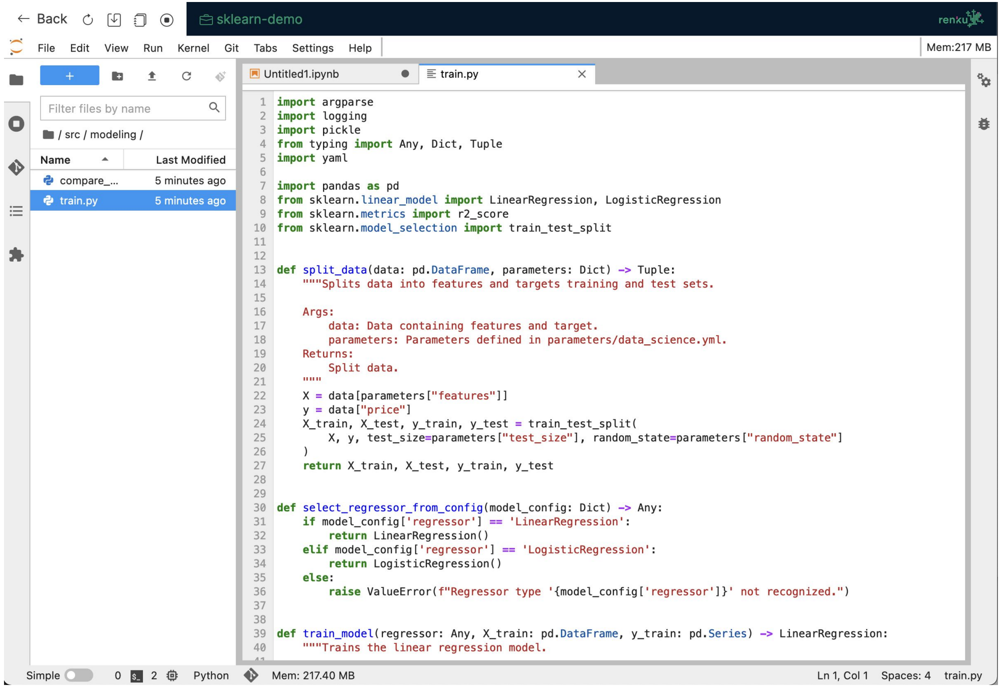
ml_extract_features_autoencoder.py

python ml_extract_features_autoencoder.py

feature_representation

Environment

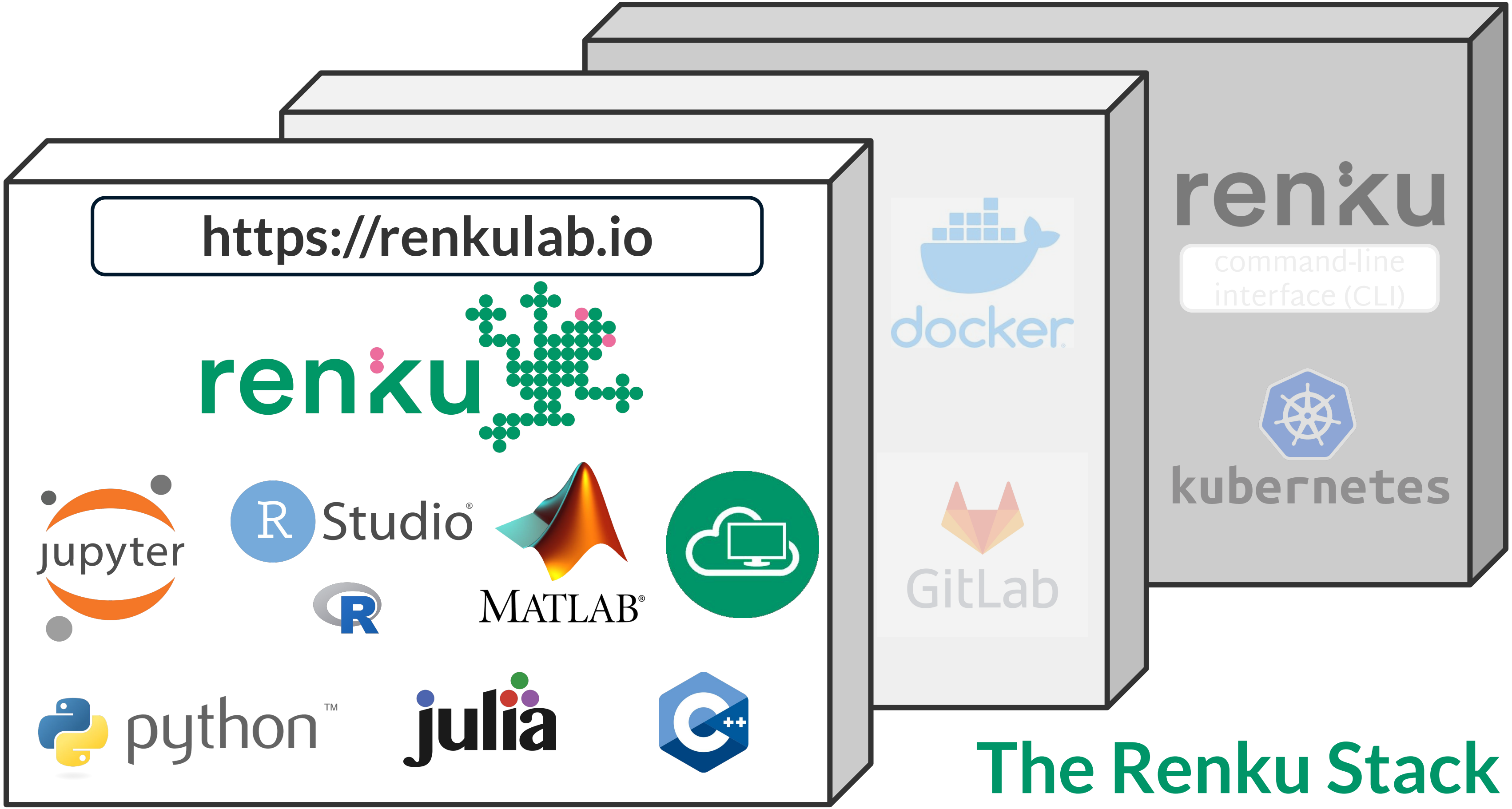
- Easy access to shared compute and storage
- Containers for reproducibility and portability, templates for consistency
- Maintained library of images to keep things up-to-date; install apps, dashboards, desktops etc.
- Configurable access to resources
- Shared project data sources



```python
import argparse
import logging
import pickle
from typing import Any, Dict, Tuple
import yaml

import pandas as pd
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split


def split_data(data: pd.DataFrame, parameters: Dict) -> Tuple:
    """Splits data into features and targets training and test sets.

    Args:
        data: Data containing features and target.
        parameters: Parameters defined in parameters/data_science.yml.
    Returns:
        Split data.
    """
    X = data[parameters["features"]]
    y = data["price"]
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=parameters["test_size"], random_state=parameters["random_state"]
    )
    return X_train, X_test, y_train, y_test


def select_regressor_from_config(model_config: Dict) -> Any:
    if model_config['regressor'] == 'LinearRegression':
        return LinearRegression()
    elif model_config['regressor'] == 'LogisticRegression':
        return LogisticRegression()
    else:
        raise ValueError(f"Regressor type '{model_config['regressor']}' not recognized.")


def train_model(regressor: Any, X_train: pd.DataFrame, y_train: pd.Series) -> LinearRegression:
    """Trains the linear regression model.
```

For the user, there is NO vendor or technology lock-in

apart from git +
docker



https://renkulab.io

The Renku Stack

Where is Renku used?

- Public instance at renkulab.io; several other smaller instances
- Primary use-cases:
    - Teaching (courses, workshops)
    - Small teams working on data analysis projects
    - Showcasing of derived datasets and results
    - Improving reusability of data products
- Used as a “connecting” piece (e.g. enabling collaborative access to data products from MMODA)

What’s ahead

- Renku as the “middle layer”/connector of code, compute and data
- More comprehensive overviews of where and how data is used
- Better integration with data providers and institutional repositories
- Our goal is to make data “alive” – how can we do better? What would you imagine to be useful for your community?
- High-level organizational, group, and topical views based on the knowledge graph

We want to hear from you!

🙋 Try out Renku

- renkulab.io - *Public*

📃 Renku Docs

❓ Run into a problem?

- Post on Discourse (our forum)
- Submit a bug report

💡 Feature Request?

- Discourse!