Sustainable data science with the Renku platform

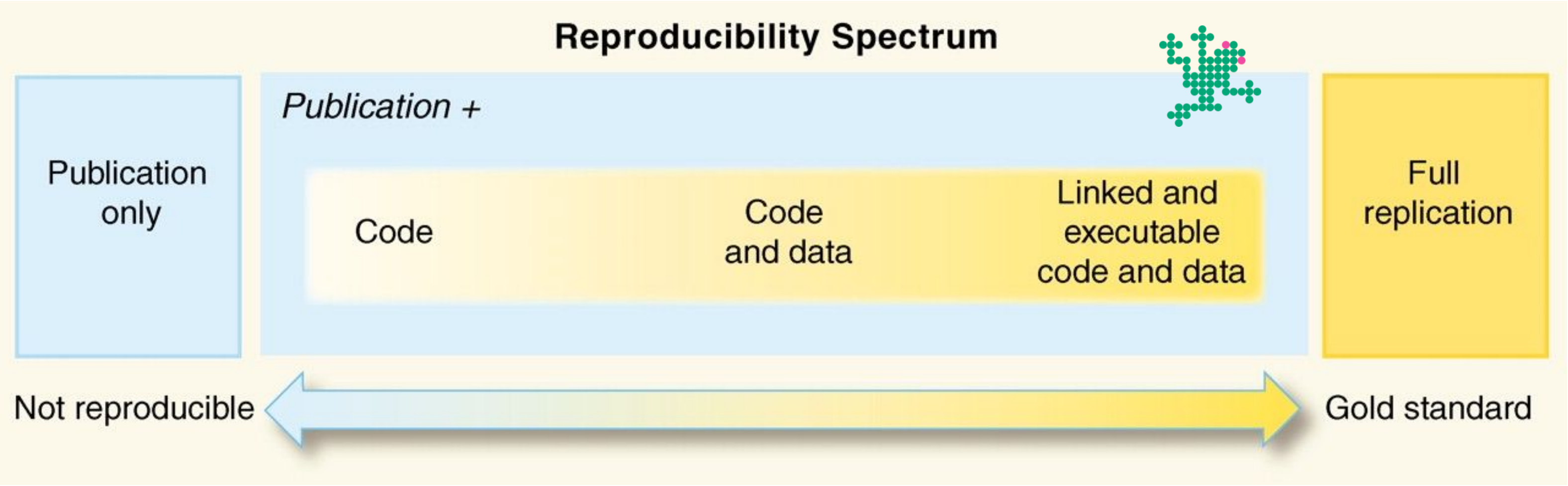
elisabet.capon@sdsc.ethz.ch

13. December 2023

In Swiss Reproducibility Network Seminar

Computational reproducibility and Renku

“ The reproducibility standard is based on the fact that every computational experiment has, in theory, a detailed log of every action taken by the computer. Making these computer codes available to others provides a level of detail regarding the analysis that is greater than the analagous noncomputational experimental descriptions printed in journals using a natural language.” Peng (2011)



Roger D. Peng, Reproducible Research in Computational Science. *Science* **334**, 1226-1227(2011). DOI:10.1126/science.1213847

Motivation

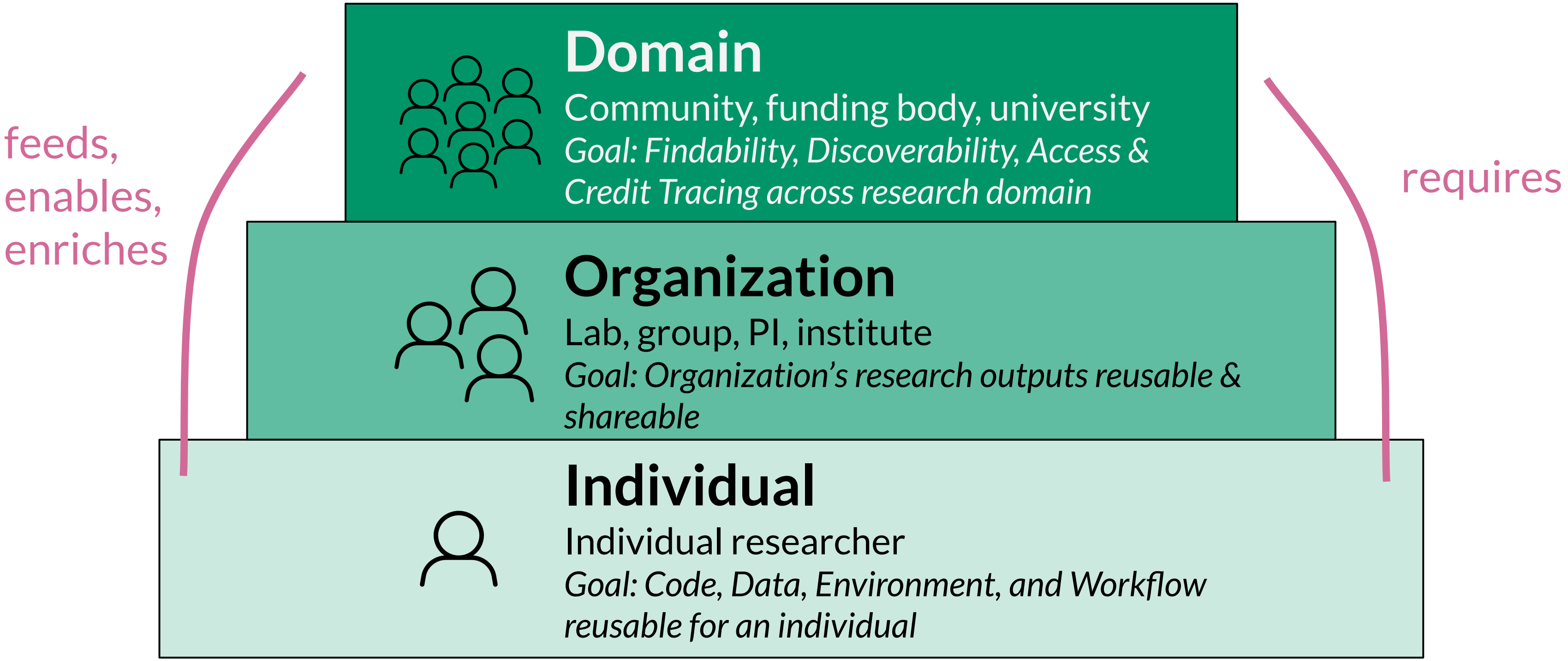
In data science,

organizations and individuals

generate **knowledge** about how to

extract value from **data.**

The Renku User Community Pyramid



feeds,
enables,
enriches

requires

Domain

Community, funding body, university

Goal: Findability, Discoverability, Access & Credit Tracing across research domain

Organization

Lab, group, PI, institute

Goal: Organization's research outputs reusable & shareable

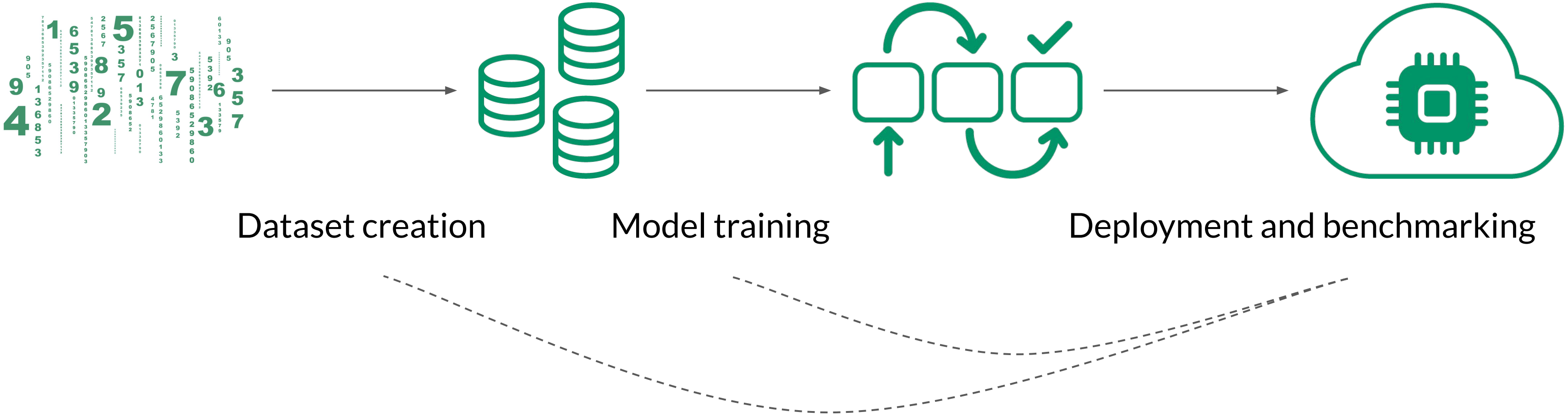
Individual

Individual researcher

Goal: Code, Data, Environment, and Workflow reusable for an individual

A sustainable system for Data Science

A sustainable ecosystem must consider the entire lifecycle



→ Goal: enable **an ecosystem** where these steps are carried out in a **self-documenting** way to **improve** individual **productivity**, **enable collaboration**, and **enrich the community**

Motivation

In data science,

organizations and individuals

generate **knowledge** about how to

extract value from **data.**

Motivation

Collaboration, teams, sharing, cross-project

In data science,

organizations and individuals

generate **knowledge** about how to

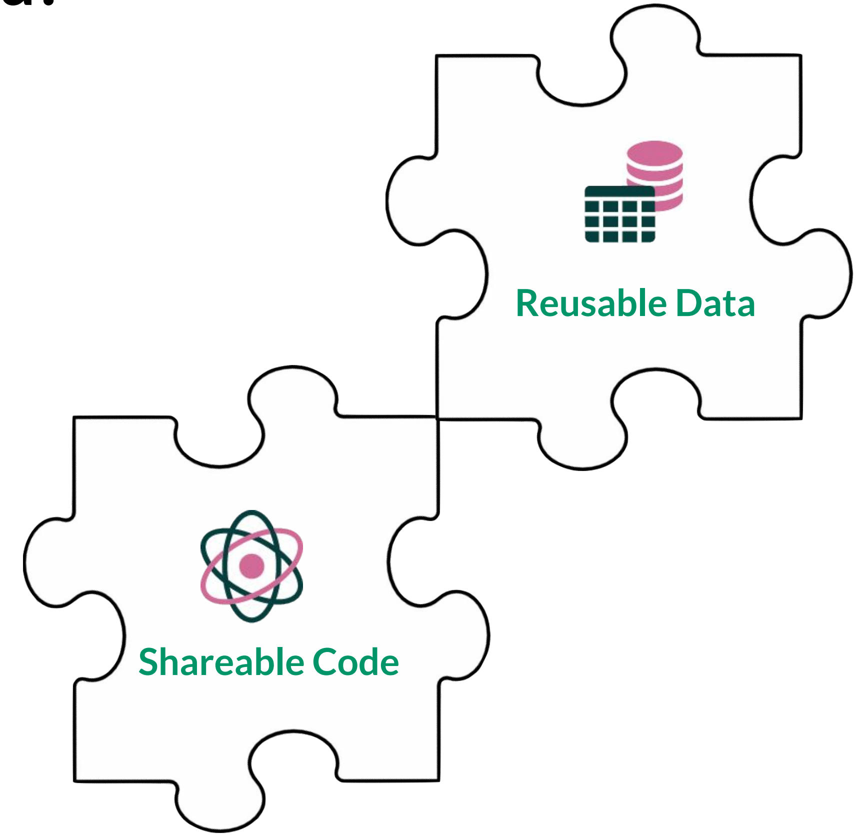
How? Why? What is related? What is connected? Who?

extract value from **data.**

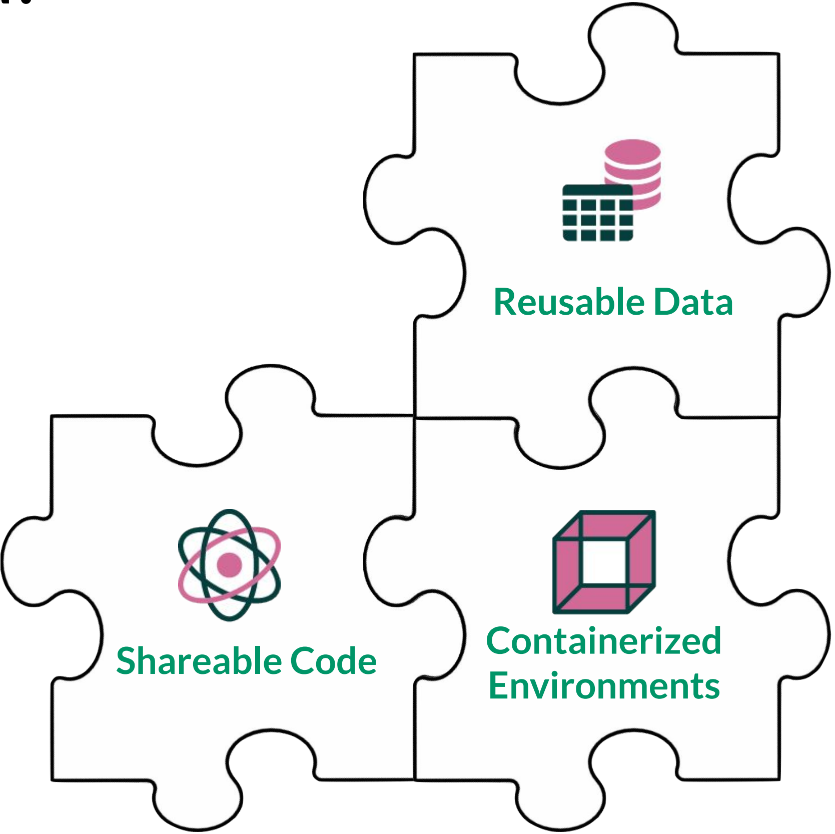
Processing, code, algorithms, pipelines, compute, repeat, follow, tracking

Files, buckets, descriptions, tags, exploration

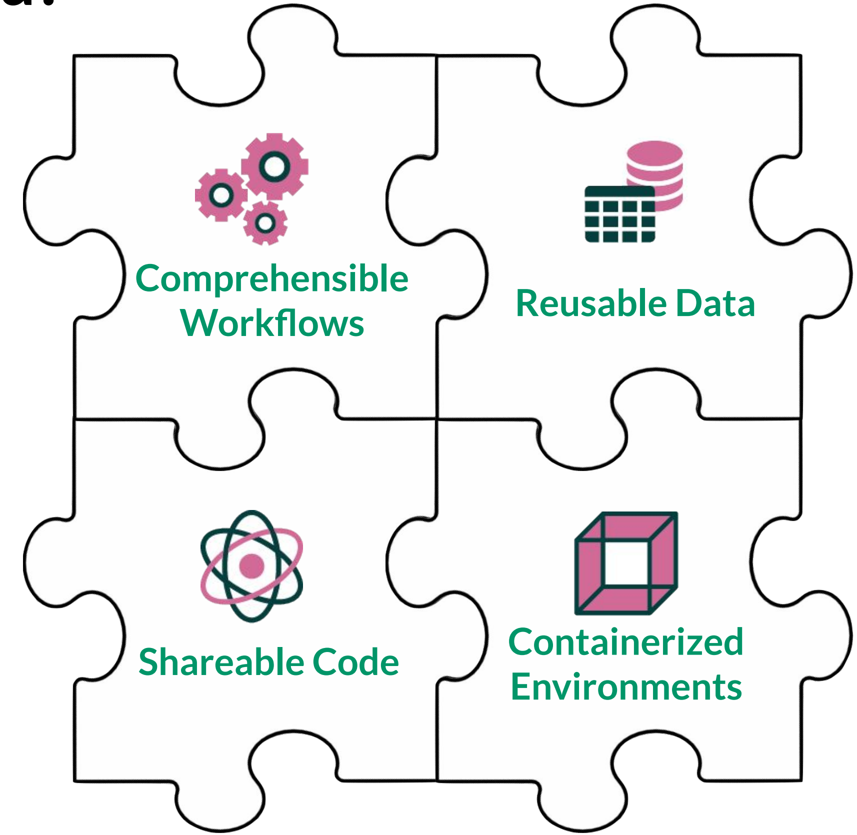
What is Renku?



What is Renku?



What is Renku?

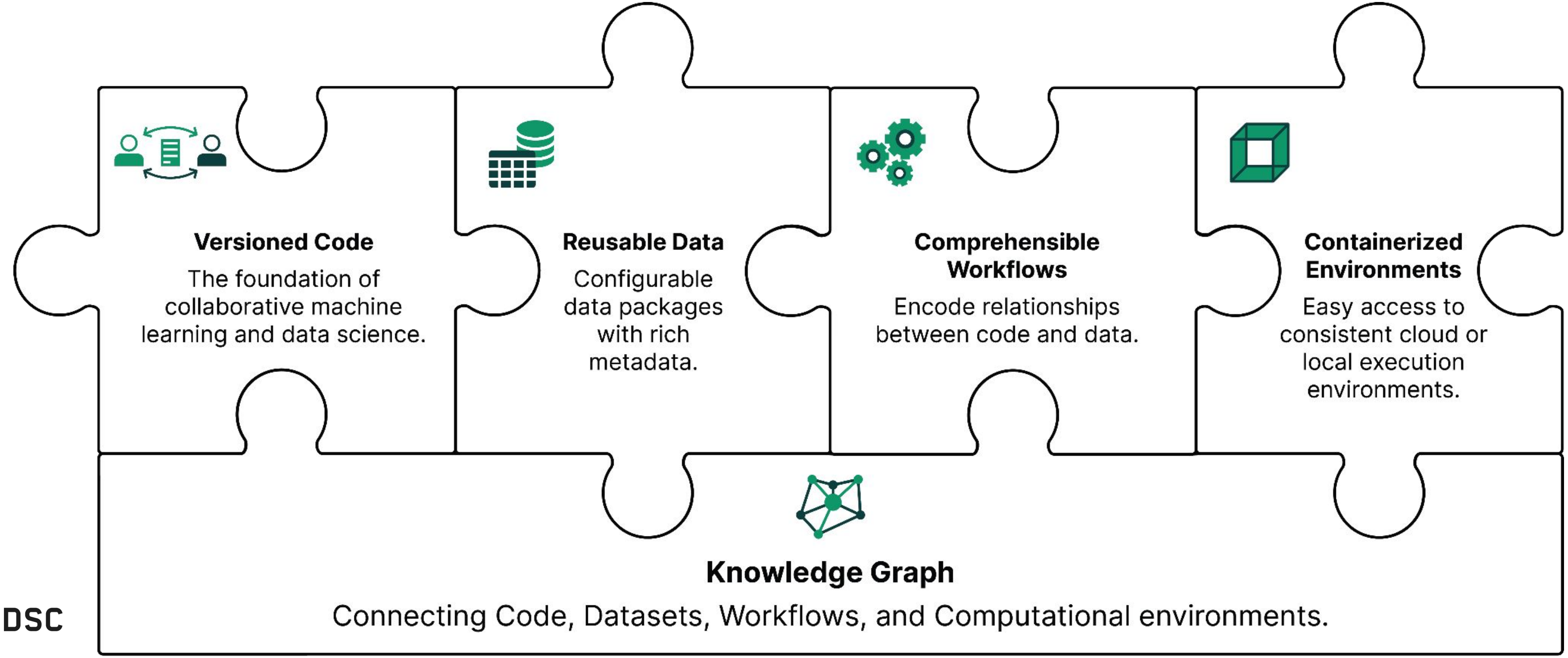


What is Renku?

Making sustainable data science practices accessible & easy

RenkuLab
Flexible & efficient compute infrastructure

Renku CLI
Renku on your own machine



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.

SDSC

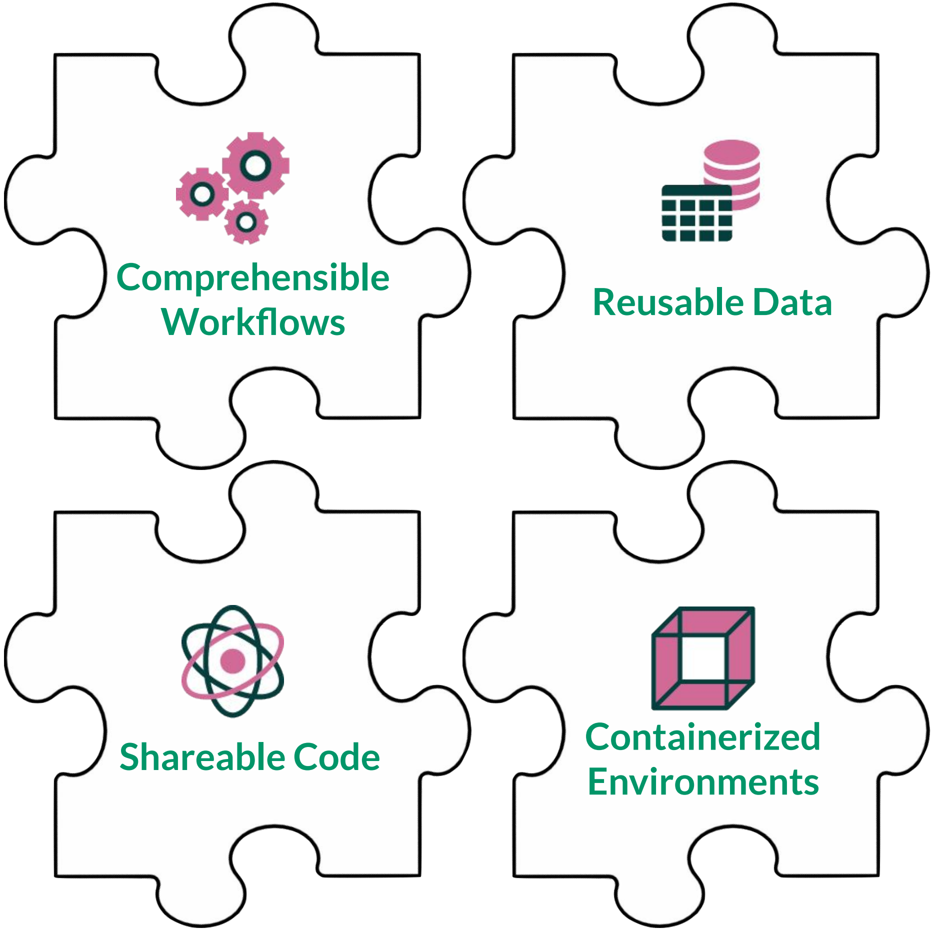
What is Renku?

current:
- Renku WF

upcoming:
- Track data use in (almost) any WF system

current:
- GitLab (renku)

upcoming:
- GitHub
- GitLab (any)



current:
- git-LFS
- S3

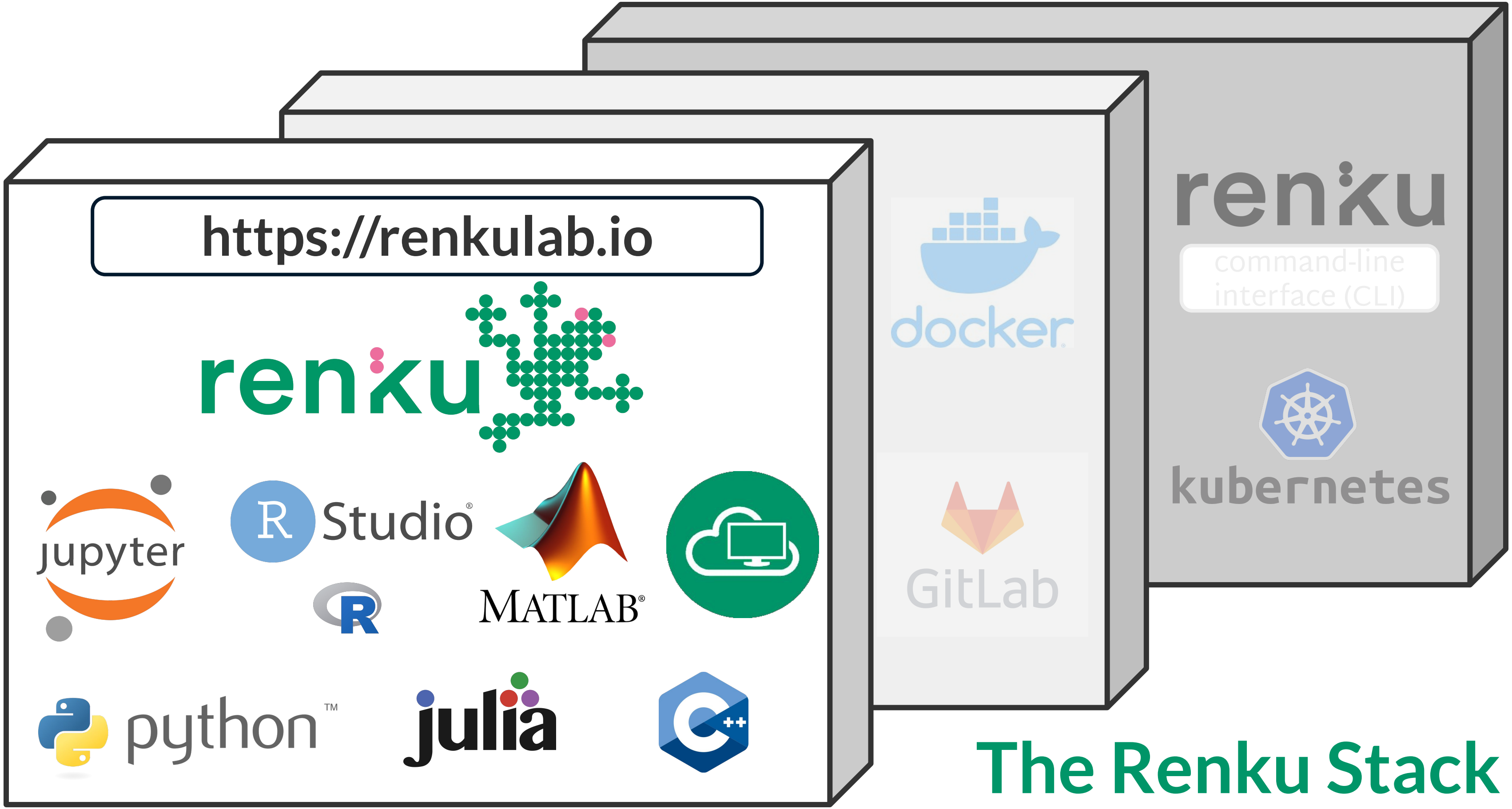
upcoming:
- WebDav
- Azure Blob
- NFS
- Domain repository connectors

current:
- Docker + gitlab-ci

upcoming:
- Flexible env. definition
- Use additional high-level standards for interoperability

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

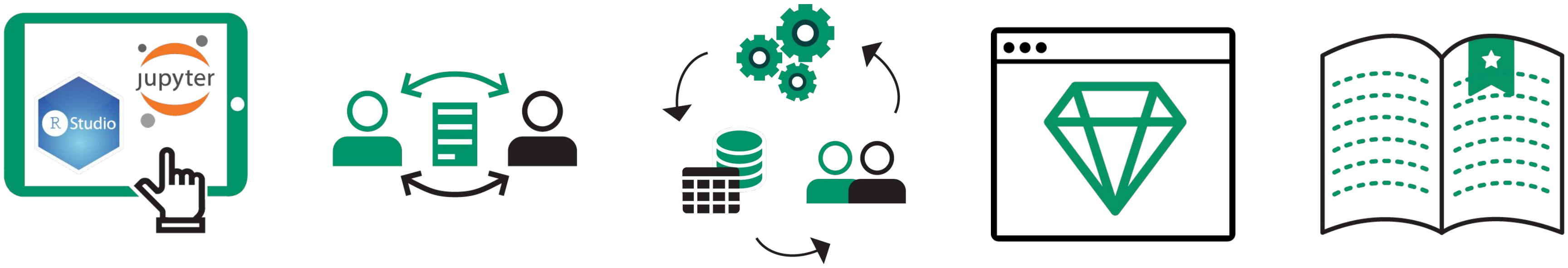
apart from git +
docker



The Renku Stack

Renku for sustainable data science

- **Build** datasets and models with **transparency** in mind from the outset
- Easily **collaborate** with others and allow them to **reuse** and **reproduce** your work
- Make results more accessible by **showcasing** their applications
- Simplify dataset and model **access**



Build Collaborate Reproduce Showcase Publish

SDSC renku

🌐 https://renkulab.io/

Demo

- https://renkulab.io/

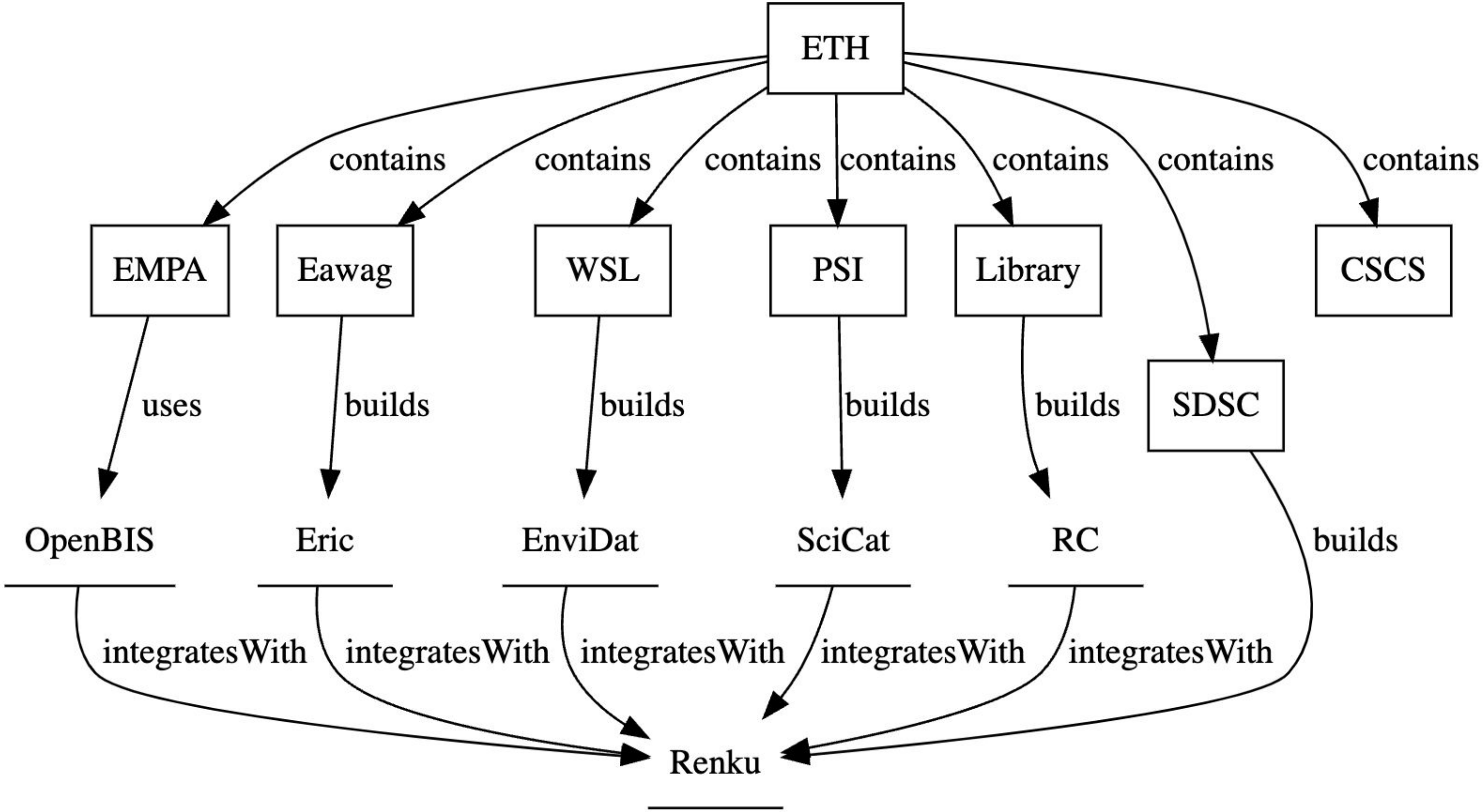
- Renku Features on a sample ML problem in Python (link)

Roadmap

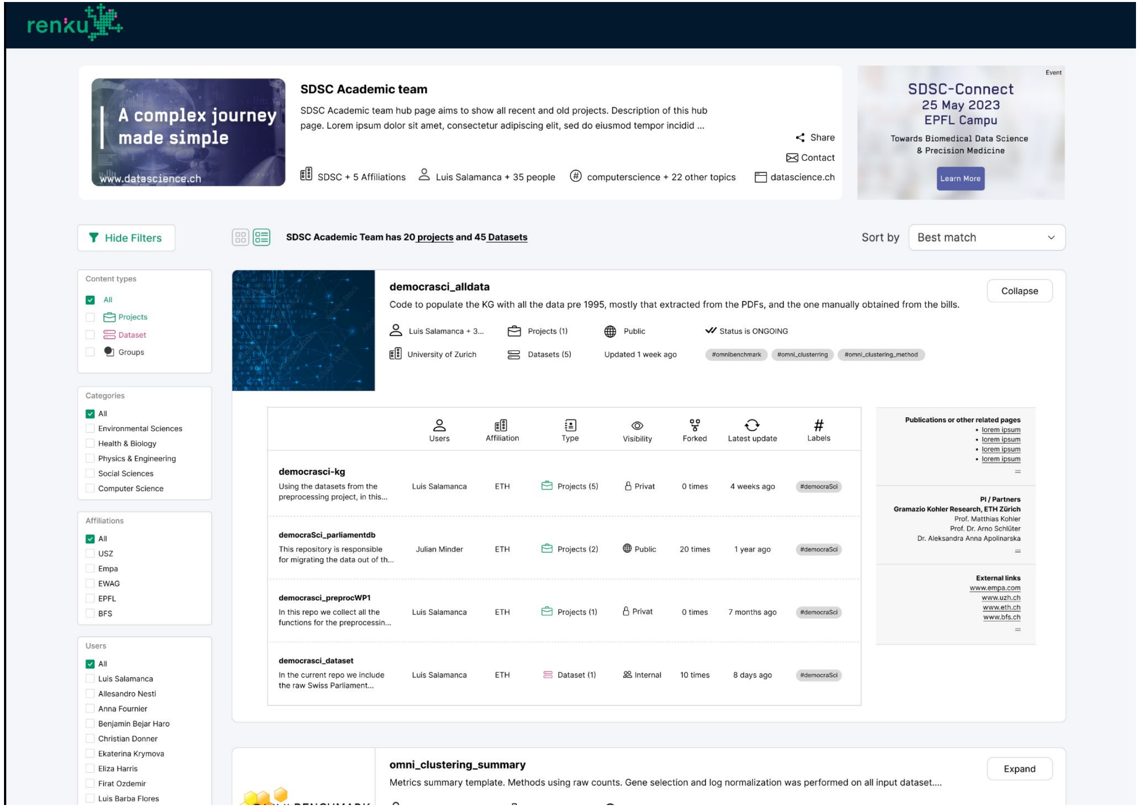
- Access to data repositories
- Hub pages
- Apps
- Renku Native Projects

SDSC renku

Data repositories - Open Research Data Initiative



Hub Pages



SDSC Academic team

SDSC Academic team hub page aims to show all recent and old projects. Description of this hub page. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incidid ...

Share

Contact

SDSC + 5 Affiliations | Luis Salamanca + 35 people | computerscience + 22 other topics | datascience.ch

Event

SDSC-Connect
25 May 2023
EPFL Campu

Towards Biomedical Data Science
& Precision Medicine

Learn More

Hide Filters

SDSC Academic Team has 20 projects and 45 Datasets

Sort by: Best match

Content types
- ☑ All
- □ Projects
- □ Dataset
- □ Groups

Categories
- ☑ All
- □ Environmental Sciences
- □ Health & Biology
- □ Physics & Engineering
- □ Social Sciences
- □ Computer Science

Affiliations
- ☑ All
- □ USZ
- □ Empa
- □ EWAG
- □ EPFL
- □ BFS

Users
- ☑ All
- □ Luis Salamanca
- □ Allesandro Nesti
- □ Anna Fournier
- □ Benjamin Bejar Haro
- □ Christian Donner
- □ Ekaterina Krymova
- □ Eliza Harris
- □ Firat Ozdemir
- □ Luis Barba Flores

democrasci_alldata

Code to populate the KG with all the data pre 1995, mostly that extracted from the PDFs, and the one manually obtained from the bills.

Collapse

Luis Salamanca + 3... | Projects (1) | Public | Status is ONGOING

University of Zurich | Datasets (5) | Updated 1 week ago | #omnibenchmark #omni_clusterring #omni_clustering_method

	Users	Affiliation	Type	Visibility	Forked	Latest update	Labels
democrasci-kg Using the datasets from the preprocessing project, in this...	Luis Salamanca	ETH	Projects (5)	Privat	0 times	4 weeks ago	#democraSci
democraSci_parliamentdb This repository is responsible for migrating the data out of th...	Julian Minder	ETH	Projects (2)	Public	20 times	1 year ago	#democraSci
democrasci_preprocWP1 In this repo we collect all the functions for the preprocessin...	Luis Salamanca	ETH	Projects (1)	Privat	0 times	7 months ago	#democraSci
democrasci_dataset In the current repo we include the raw Swiss Parliament...	Luis Salamanca	ETH	Dataset (1)	Internal	10 times	8 days ago	#democraSci

Publications or other related pages
- lorem ipsum
- lorem ipsum
- lorem ipsum
- lorem ipsum

...

PI / Partners
Gramazio Kohler Research, ETH Zürich
Prof. Matthias Kohler
Prof. Dr. Arno Schlüter
Dr. Aleksandra Anna Apolinarska
...

External links
www.empa.com
www.uzh.ch
www.eth.ch
www.bfs.ch
...

omni_clustering_summary

Metrics summary template. Methods using raw counts. Gene selection and log normalization was performed on all input dataset....

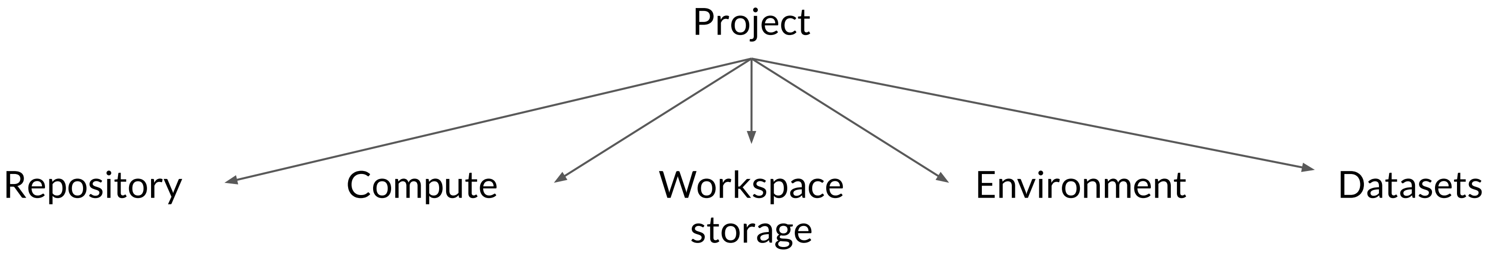
Expand

Apps - Motivation

1. Apps are hard to create on RenkuLab.
    a. The templates don’t work, we don’t have templates for RShiny, etc.
2. I want to share my app with the public, but not the project or session resources.
3. I don’t want to go through the Jupyter navigation to go to the app.
    a. This is possible right now, but not well advertised.
4. I want my app to launch/open quickly (hosted, not launched in session)

Renku Native Projects - Motivation

- Independence from GitLab
- Each of the elements of the projects stand for an independent resource
- More modular set-up
- Separate Datasets from Git
- Simplify the procedure on onboarding



Project

Repository

Compute

Workspace storage

Environment

Datasets

Note: current Renku projects != Renku-native projects

Definition of Renku Native building blocks

Repository (version control): Github or Gitlab repository

Compute: Currently and in the near future only our K8s cluster, but we could add more options here (i.e. AWS, Azure, etc.)

Workspace storage: Read/write storage backed by K8s PVC, S3, Azure Blob, etc.

Environment: Docker image

Datasets: Read-only storage containing published data stored in S3, Azure Blob, etc.

We want to hear from you!

https://docs.google.com/forms/d/e/1FAIpQLSc4qB5KUX5cwIP2f1nNX1BA1zt-NYt28sIPxXZL6qjeMtpheQ/viewform?usp=pp_url

🙋 **Try out Renku**

- renkulab.io - *Public*

📃 **Renku Docs**

❓ **Run into a problem?**

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

💡 **Feature request or idea?**

- Check out our roadmap!
- Ask a question or propose a feature

👷 **Contribute to Renku**

- Renku Design Docs

Computational reproducibility in Renku

Public projects in Python:

- "Increased dose efficiency of breast CT with grating interferometry", Rawlik, Michał; Pereira, Alexandre; Spindler, Simon; Wang, Zhentian; Romano, Lucia; Jefimovs, Konstantins; Shi, Zhitian; Polikarpov, Maxim; Xu, Jinqiu; Zdora, Marie-Christine; van Gogh, Stefano; Stauber, Martin; Yukihara, Eduardo G; Christensen, Jeppe B; Kubik-Huch, Rahel A; Niemann, Tilo; Leo, Cornelia; Varga, Zsuzsanna; Boss, Andreas; Stampanoni, Marco (2023). Optica, 10(7):938.

Public projects in R:

- https://renkulab.io/projects/babraham-reproducibilitea-journal-club/babraham-reproducibilitea-journal-club
- https://renkulab.io/projects/hdbi/data-management/hdbi-data-resource
- “Lütge A, Zyprych-Walczak J, Brykczynska Kunzmann U, Crowell HL, Calini D, Malhotra D, Soneson C, Robinson MD. CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. Life Sci Alliance. 2021 Mar 23;4(6):e202001004. doi: 10.26508/lsa.202001004. PMID: 33758076; PMCID: PMC7994321” (link)