



RENKU - 連句

Making Data Science open and reproducible

Rok Roškar and the SDSC Renku team

About me

- Rok Roškar, PhD in computational astrophysics
- Spent the past ~4 years grappling with “big” data problems in academia @ Scientific IT Services, ETH
- Currently a data scientist/software engineer and serving as the deputy Chief Architect for our platform
- You can contact me at rok.roskar@sdsc.ethz.ch or @rokstars

Swiss Data Science Center (SDSC)

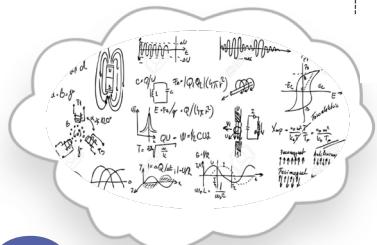
Foster adoption of data science both in academia and industry



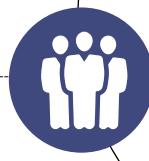
How can I best match
the right drug with the
right dosage to the right
patient at the right time?



What is the hyperplane
that best separates two
classes of points in
multidimensional space?



Domain experts



Data scientists



Data providers

ETH zürich +

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

How is my data protected?
How private is it?
How exactly is it used?



FAQs in Data-Driven Research

complexity

- Where did this model come from?
- How did my student's/coworker's/boss' code run?
- How does this new data change our predictions?
- Can I use your data? With your code? On your computer? In the cloud?
- Has anyone ever trained an <XYZ-algorithm> on this data?
- Who is using my data? and my algorithm? Why are they not citing me?!
- How do I run my analysis on your confidential data?

Benefits of best (open) practices:

If you can answer these questions confidently:

- Your results are trustworthy
- You can collaborate easily
- Your team is efficient
- You participate in Open Science/Open Data
- You are properly acknowledged (and you acknowledge others!)
- ... and you can be held accountable

Context: Open + FAIR

- Open Science, Open Data
 - Public repositories e.g. Zenodo, library archives
 - Great tools like Invenio from CERN or CKAN
 - Required by many (most?) public funding bodies
- FAIR – Findable, Accessible, Interoperable, Reusable
- Great in principle, but few know how to implement?
- Disconnect between policy and work “on the ground” – need for better, more efficient tools close to the research process
- Need to provide practical benefits for using “best-practices” of data sharing/reproducibility/openness

Goals of Renku

1.

Provide the means to create
reproducible data science

2.

Facilitate the **sharing** and
reuse of research artefacts

3.

Foster a **collaborative**
environment for interactive
prototyping

4.

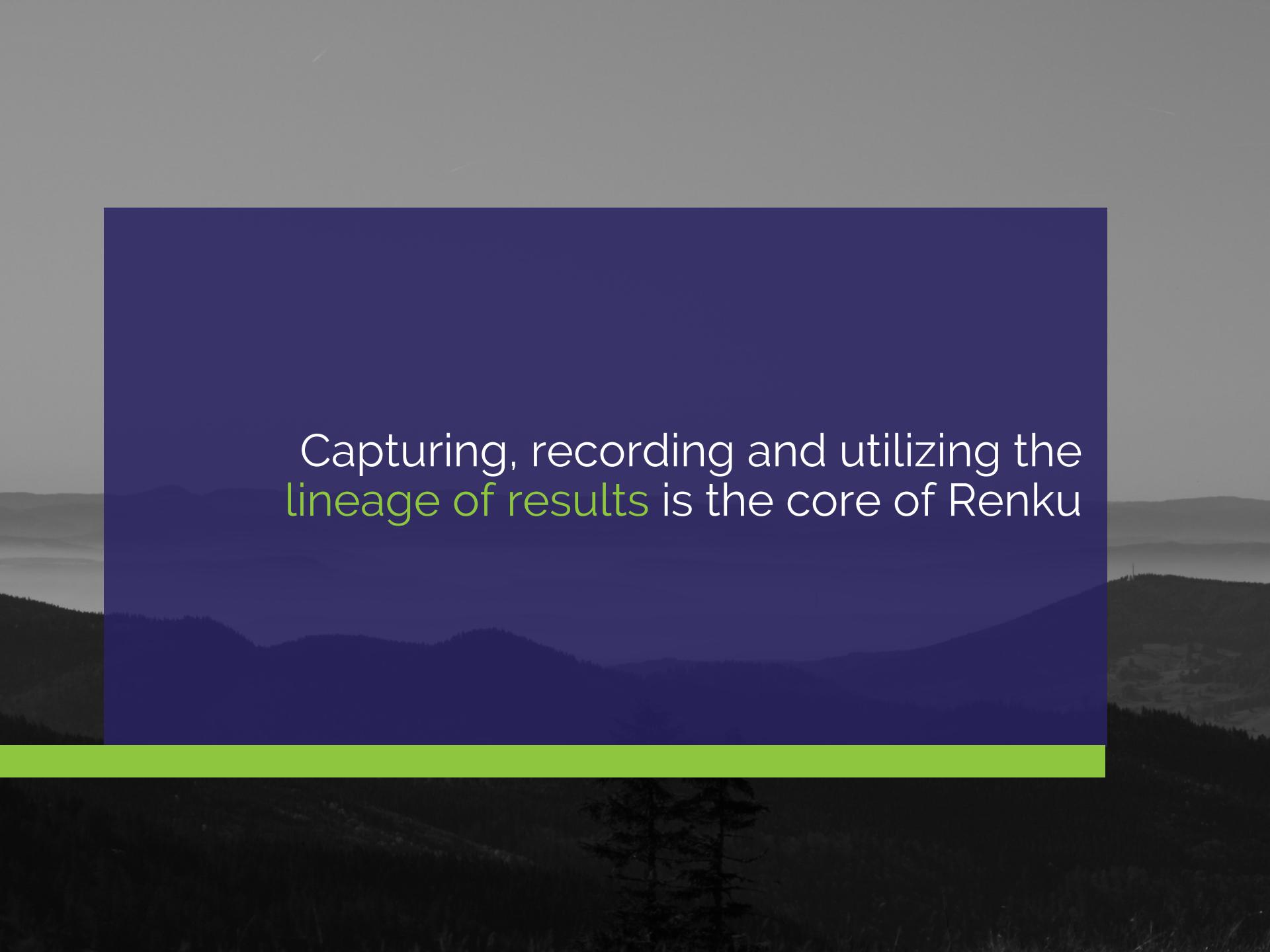
Enable the **discovery** of
relevant data and methods

5.

Allow **federated access** across institutions giving each the
freedom to impose its own access controls over resources

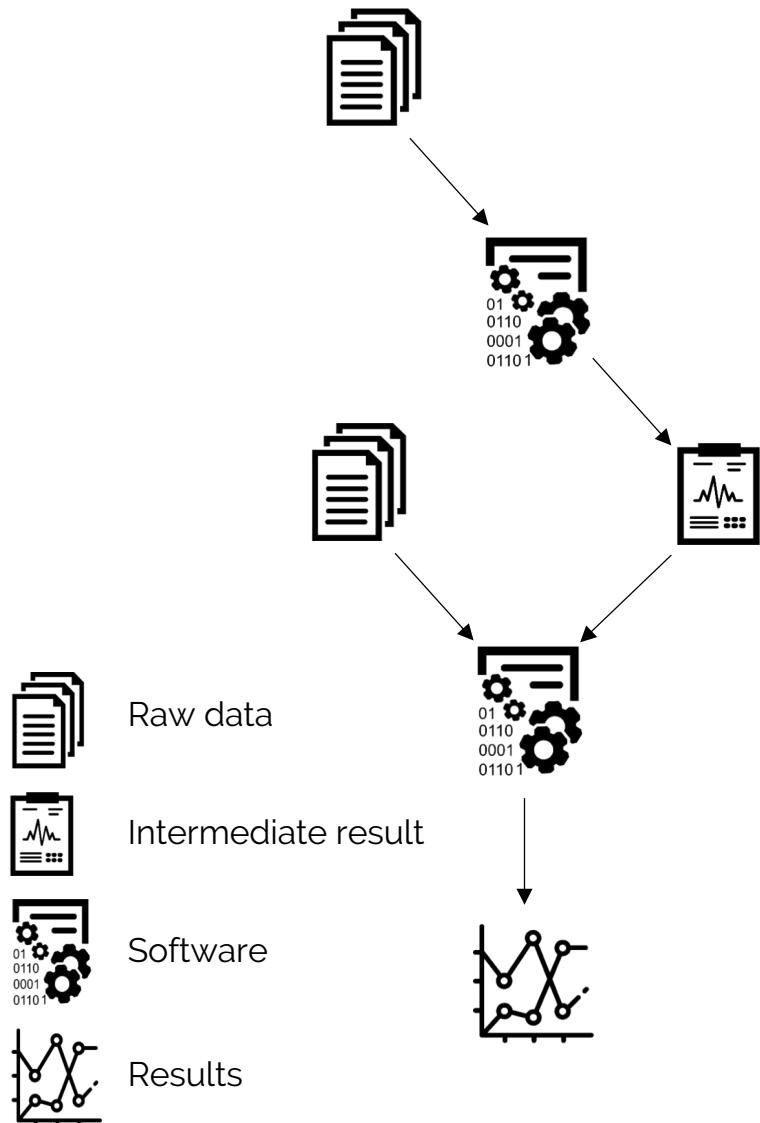
Terminology

- We borrow the **Renku** name from the Japanese word for *linked-verse poetry*
- A “**ku**” is a verse in a renku poem
- We use “**ku**” to mean a piece of the data analysis process – includes discussion, code, and results



Capturing, recording and utilizing the
lineage of results is the core of Renku

Capture the scientific process



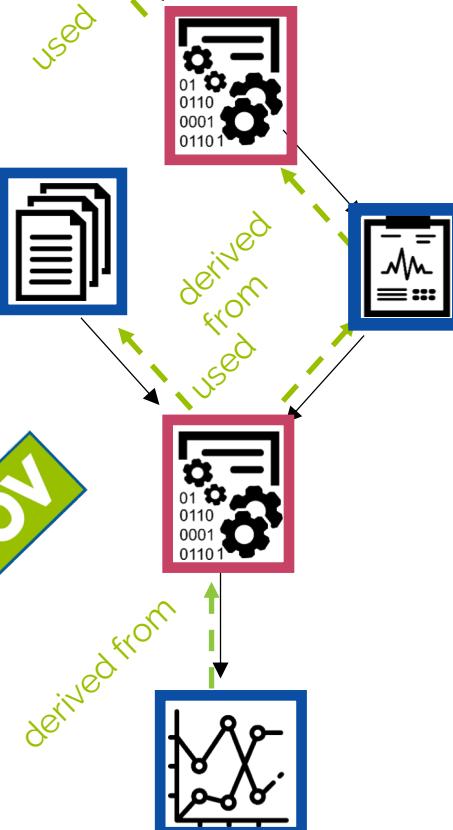
1. Inputs and outputs of analysis steps are recorded into a **knowledge graph** as the work is being done
2. Steps can be **repeated** or integrated into more complex **workflows**
3. **Provenance** of all data products is always accessible via simple tools
4. **Version control** is built-in for data, code, and workflows

Encapsulate with rich metadata



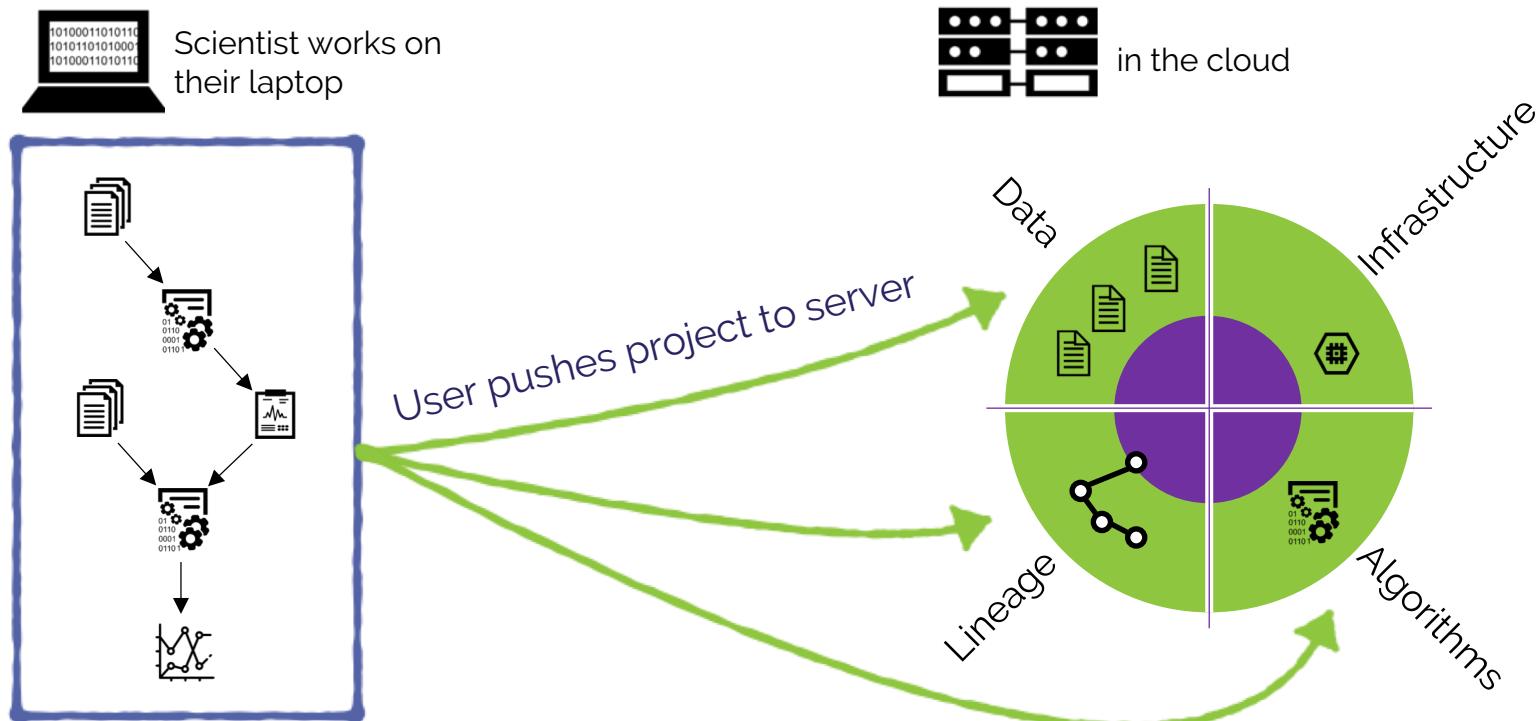
COMMON
WORKFLOW
LANGUAGE

- Metadata use Dublin Core, FOAF, and Schema.org
- Provenance graph is based on PROV-O W3C recommendation



- CWL for representing all computational steps
- Capture individual steps from user input
- Tools for constructing workflows from basic pieces
- Rely on container technologies to ensure reproducibility

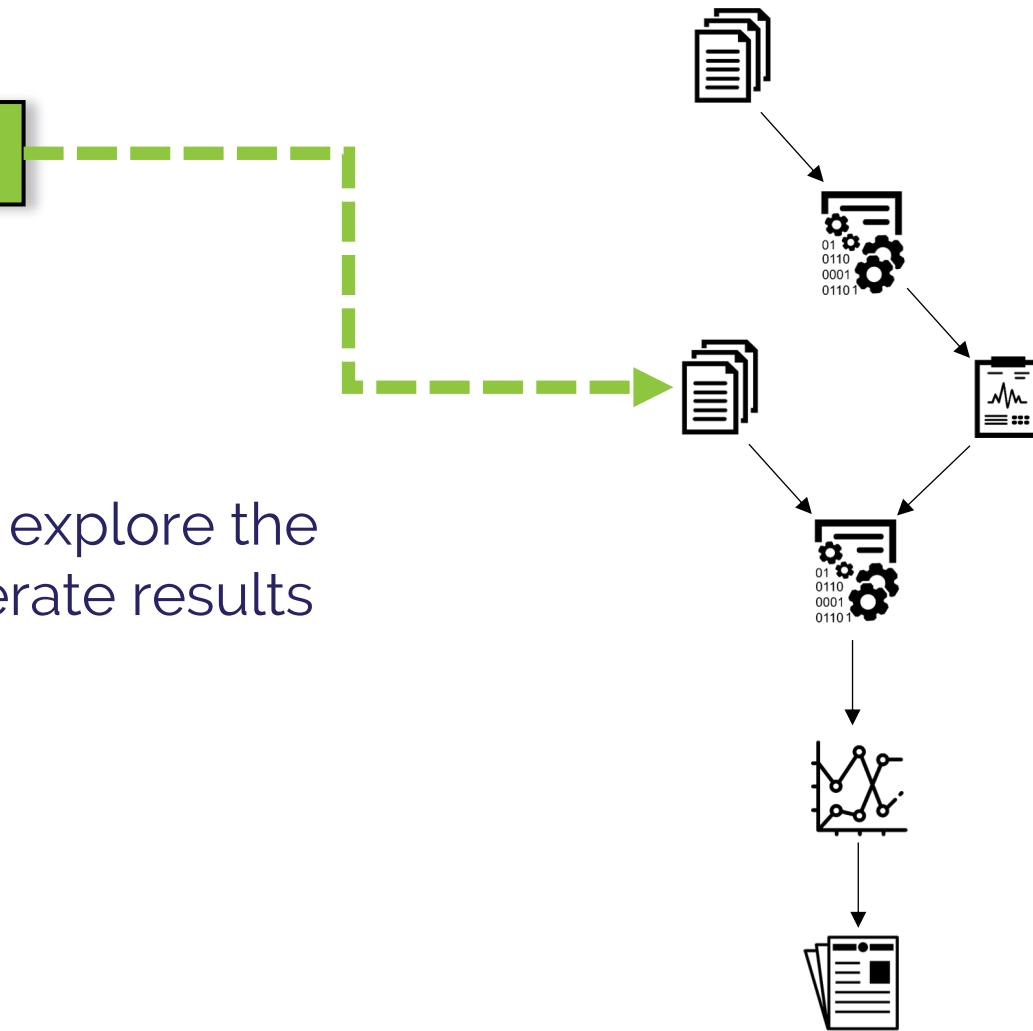
Verify and share results



1. Results are automatically verified
2. Data lineage is captured “live”
3. Knowledge graph is populated
4. A shareable interactive environment created

Discover and understand the work of others

Graph-based search...



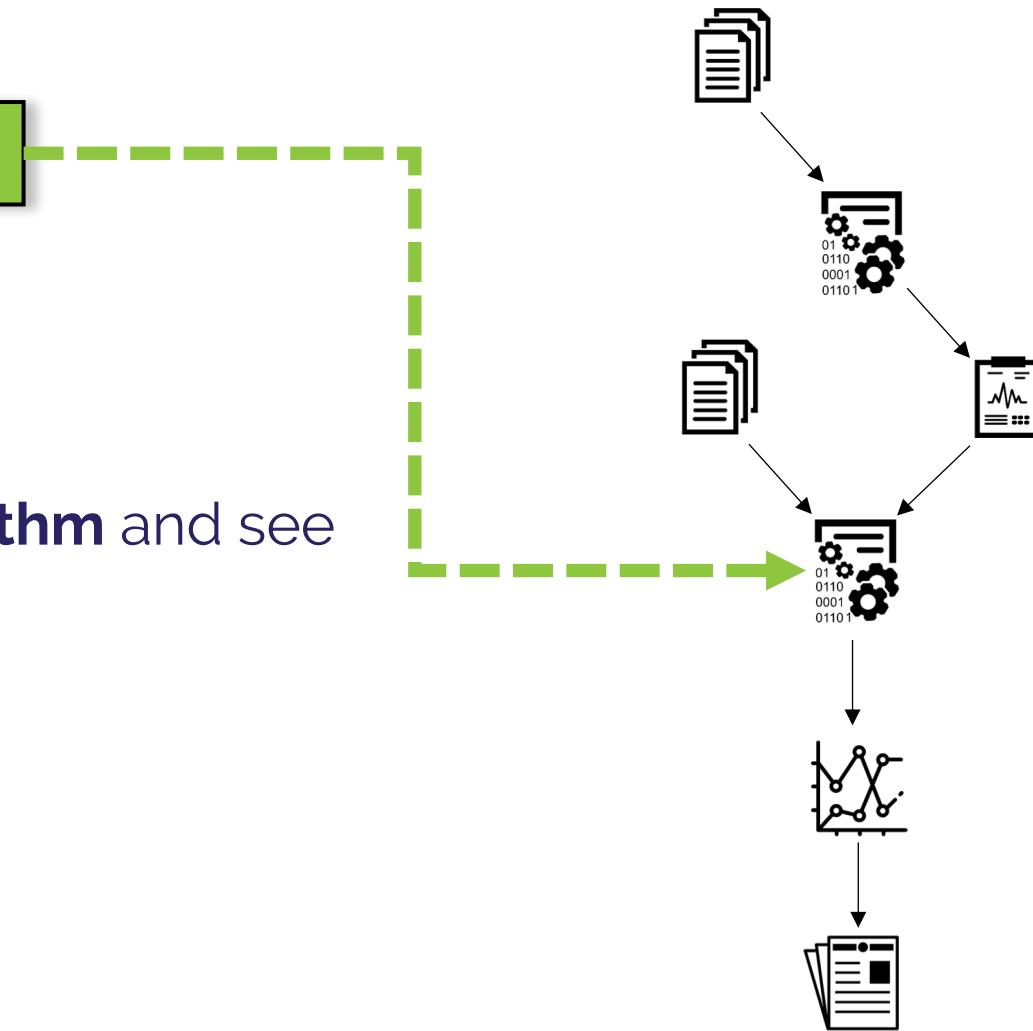
Ex: search for **data** and explore the tools to efficiently generate results

Discover and understand the work of others

Graph-based search...



Ex: search for an **algorithm** and see its applications

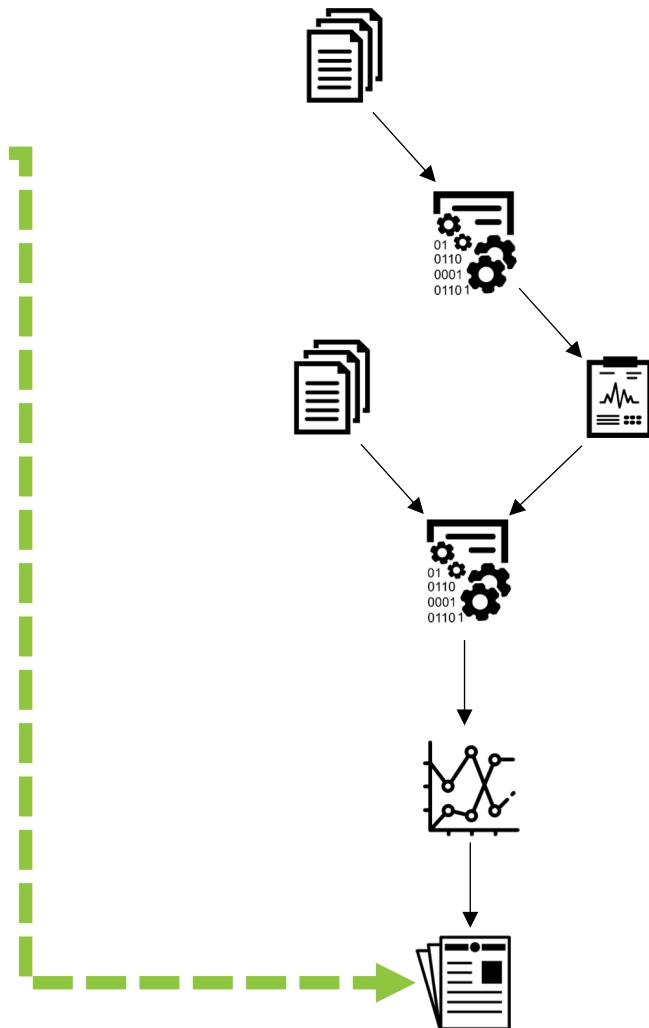


Discover and understand the work of others

Graph-based search...



Ex: search for a **publication**, obtain a full view of how the results were obtained

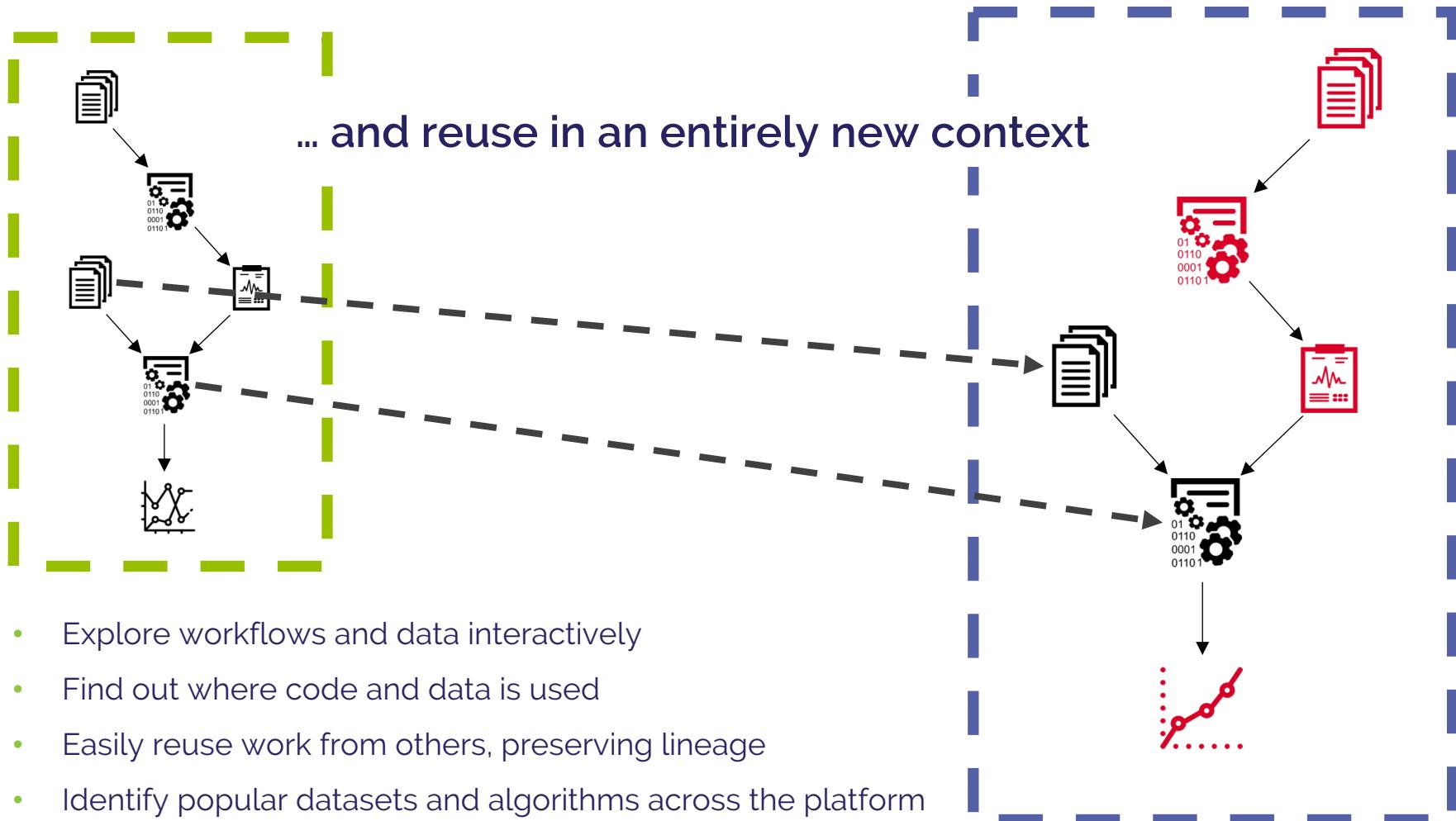


Reuse and repeat

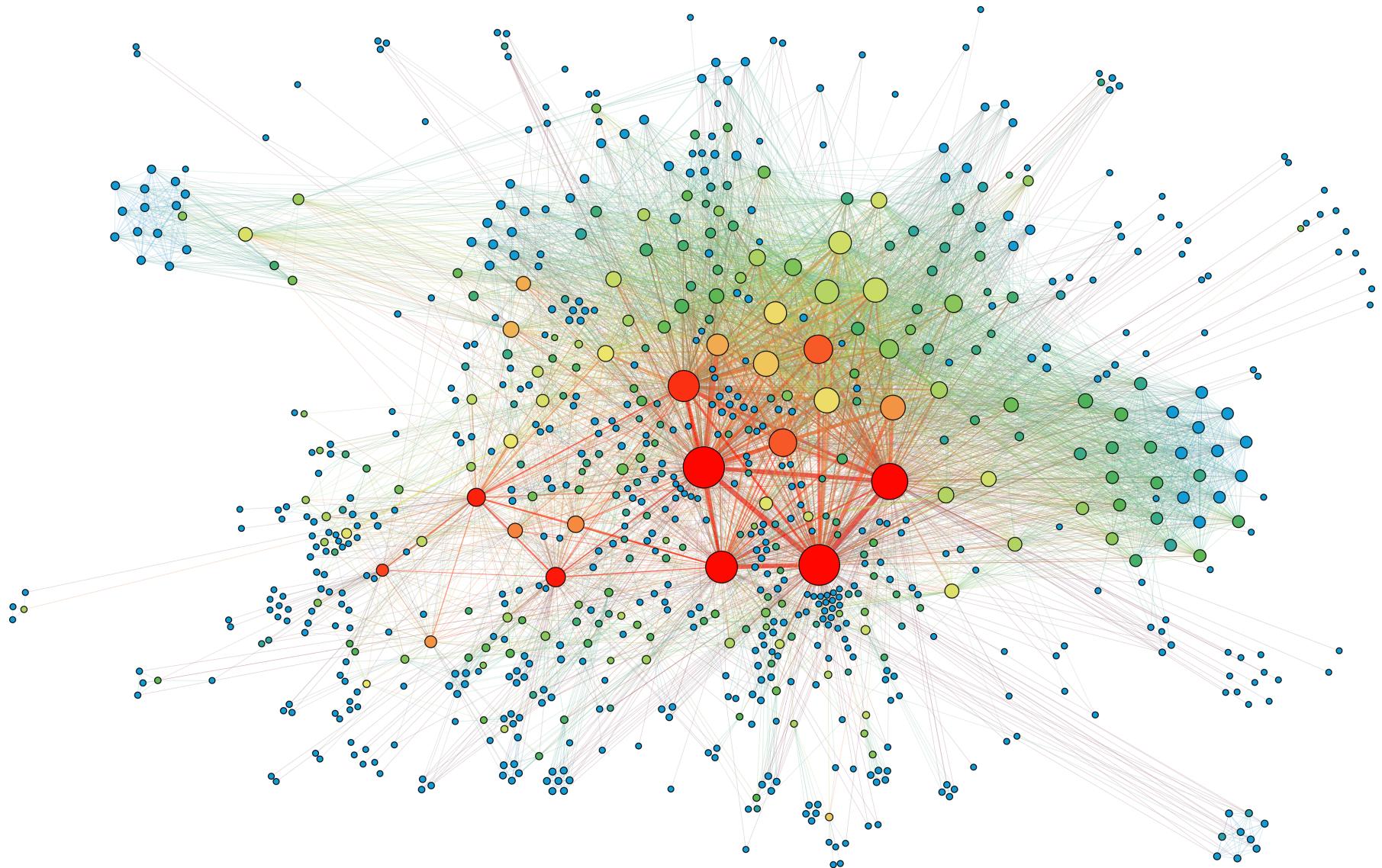
Search for relevant data or algorithms...



... and reuse in an entirely new context



A Knowledge Graph emerges



A peek into the data science process

- Who is using the data and how?
- Which algorithms are used to answer which questions?
- How to regenerate results if new data becomes available? If old data is now off-limits?
- Who to credit?
- How popular is my work/the work of my lab/my unit?

TRUST

Building easy-to-use tools on top of trusted technologies

- Renku consolidates the open-source data science and software engineering technologies into a single platform
- The user interface gently nudges users to follow best-practices and create reproducible work

```
Requirement already satisfied: html5lib<1.0b1,!>1.0b2,!>1.0b3,!>1.0b4,!>1.0b5,!>1.0b6,!>1.0b7,!>1.0b8,>=0.9999999pre in /opt/conda/lib/python3.6/site-packages (from bleach->nbconvert->jupyter->weather-ch<=0.1.0)
Requirement already satisfied: parse<=0.1.1 in /opt/conda/lib/python3.6/site-packages (from jedi>0.10->ipython>=4.0.0->ipykernel->jupyter->weather-ch<=0.1.0)
Requirement already satisfied: ptyprocess<=0.5 in /opt/conda/lib/python3.6/site-packages (from pexpect; sys_platform != "win32">ipython>=4.0.0->ipykernel->jupyter->weather-ch<=0.1.0)
Requirement already satisfied: webencodings in /opt/conda/lib/python3.6/site-packages (from html5lib<1.0b1,!>1.0b2,!>1.0b3,!>1.0b4,!>1.0b5,!>1.0b6,!>1.0b7,!>1.0b8,>=0.9999999pre>nbconvert->jupyter->weather-ch<=0.1.0)
Installing collected packages: seaborn, patsy, statsmodels, weather-ch
Running setup.py install for seaborn: started
  Running setup.py install for patsy: finished with status 'done'
  Running setup.py develop for weather-ch
Successfully installed patsy-0.5.0 seaborn-0.8.1 statsmodels-0.8.0 weather-ch
Removing intermediate container 779bbdbd76d8
--> 39778a995087
Step 10/10 : RUN rm -rf /tmp/* /root/*
--> Running in 33ae022ee08
Removing intermediate container 33ae022ee08
--> 2eb1832cded9
Successfully built 2eb1832cded9
Successfully tagged gitlab.renka.build:5081/rök/weather-ch/review-master-phenv01:25046583c71ca7b8363328e039bf6a8128b218ef
Renka tagged rök/reviews ->/projects-presentation/weather-ch renka status
On branch master
All files were generated from the latest inputs
# On branch master
#   * [pi] renga-demo ->/projects-presentation/weather-ch  renka log
#     gitignore      .gitignore
#     data/          gitlab_checkpoint/.renka.lock
#     notebooks/    requirements.txt
#     src/          Dockerfile
#     README.md
#     Tok  # master [pi] renga-demo ->/projects-presentation/weather-ch  renka log data/zh/
#     homog_mo_SMA.txt metadata.yml      standardized.csv
#     homog_mo_SMA.txt metadata.yml      standardized.csv
#     66fcff01 data/zh/standardized.csv
#     66fcff01 renka/workflow/97419ac93b5466bb3d6fc70f6ffff5ffc.python.cwl
#     54af536d data/zh/homog_mo_SMA.txt
#     Tok  # master [pi] renga-demo ->/projects-presentation/weather-ch > |
```

Command-line interface

The screenshot shows the Renku web interface for the 'weather-ch' project. At the top, there's a search bar and a navigation bar with tabs for 'Overview', 'Run', 'Files', and 'Settings'. Below the tabs, there's a section titled 'Analyze data' which says 'An investigation into weather trends in Zürich, Switzerland.' and 'Updated 1 hour ago'. Underneath it is a 'Preprocess data' section with a note 'Convert values to deviation from monthly mean' and a timestamp '5 hours ago'. A user profile for 'Rok Roskar' is shown with the message 'Updated 1 hour ago. For the pre-processing results, see this notebook'. A code editor window displays a Python script for data processing:

```
import pandas as pd
import numpy as np
import scipy
import weather_ch

import matplotlib as plt
from matplotlib import cm
matplotlib_inline()
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display, HTML
sns.set()

df = weather_ch.read_standardized('../data/zh/standardized.csv')
df.head()
```

Web-based front-end

First steps – on the laptop

```
$ renku init
$ renku dataset create zh
$ renku dataset add http://www.meteoschweiz.admin.ch/...
$ tree
.
└── data
    └── zh
        ├── homog_mo_SMA.txt
        └── metadata.yml
```

First steps – on the laptop

```
$ renku run papermill notebooks/GettingStarted.ipynb
$ renku run papermill notebooks/PreprocessData.ipynb
$ renku log

* 9f3cc772 figs/temperature.png
|
*   9f3cc772 .renku/workflow/6f0e4ff58b5d41588eea1dc16aaa4a29_papermill.cwl
| \
@ | c2b077f2 notebooks/PreprocessData.ipynb
/
|
* 847a3bb1 data/zh/standardized.csv
|
*   847a3bb1 .renku/workflow/7596966fe4d6454da450f22079b847ca_papermill.cwl
| \
@ | f35542f6 notebooks/GettingStarted.ipynb
/
@ 5d759ae5 data/zh/homog_mo_SMA.txt
```

Lineage is captured and recorded in the local knowledge graph

First steps – on the laptop

```
$ <modify inputs>  
$ renku status  
On branch master  
Files generated from newer inputs:  
(use "renku log [<file>...]" to see the full lineage)  
(use "renku update [<file>...]" to generate the file from its latest inputs)  
  
    data/zh/standardized.csv: data/zh/homog_mo_SMA.txt#5d759ae5  
    figs/temperature.png: data/zh/homog_mo_SMA.txt#5d759ae5  
    notebooks/GettingStarted-run.ipynb: data/zh/homog_mo_SMA.txt#5d759ae5  
    notebooks/PreprocessData-run.ipynb: data/zh/homog_mo_SMA.txt#5d759ae5  
  
Input files used in different versions:  
(use "renku log --revision <sh1> <file>" to see a lineage for the given  
revision)  
  
    data/zh/homog_mo_SMA.txt: 520c1347, 5d759ae5
```

Outdated outputs are detected and can be regenerated automatically

Continuing on the web

Search RENGA  Projects  

weather-ch

An investigation into weather trends in Zürich, Switzerland.

Updated 14 minutes ago.

Overview Kus Files Settings

open  **Data analysis**

Perform analysis of the data to understand the weather trends. Updated 1 minute ago.

complete  **Preprocessing data**

Could we try to convert values to deviation from monthly mean? I think those would be more meaningful. Updated 2 minutes ago.

complete  **Reading input data**

Could you show me how to read in this data? Updated 2 minutes ago.

linear models **python** **weather**  unstar 2

Data analysis

 **Rok Roškar** Updated 1 minute ago.
A preliminary analysis of the data can be seen [here](#)

```
import pandas as pd
import numpy as np
import scipy
import os
import os.path

from matplotlib import cm
from IPython.display import display, HTML
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt

import seaborn as sns
```

Launch Notebook 

Upcoming work

- Begin mining the Knowledge Graph
- Continue enhancing the UI: workflows, graph interactivity, execution
- Enable execution of scalable workflows "in the cloud" and on HPC resources
- Fine-grained access controls everywhere (as open as possible, as closed as necessary)
- Federation: one interface, many resources

Current status

Platform is under very **active** development:

<https://github.com/SwissDataScienceCenter>

renku — services & deployment recipes

renku-python — CLI and Python API

Contact us if you are interested!
(also, we're hiring)

