ETH-RAT

SDSC

RENKU - 連句

Making Data Science open and reproducible

Rok Roškar and the SDSC Renku team

opendata.ch/2018 | 03.07.2018

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)

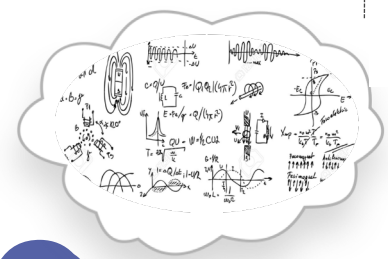
Foster adoption of data science both in academia and industry

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

ETH zürich + EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Domain experts

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



Data scientists

Data providers

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

SDSC

FAQs in Data-Driven Research

- 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

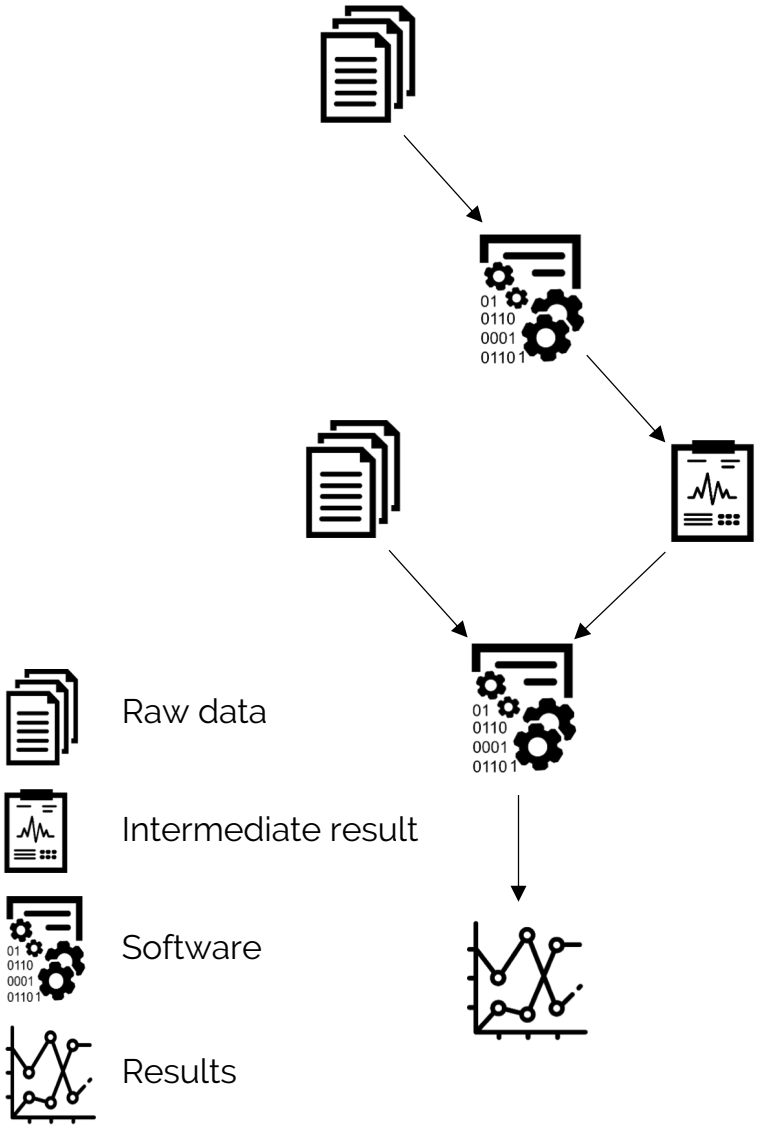
SDSC

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

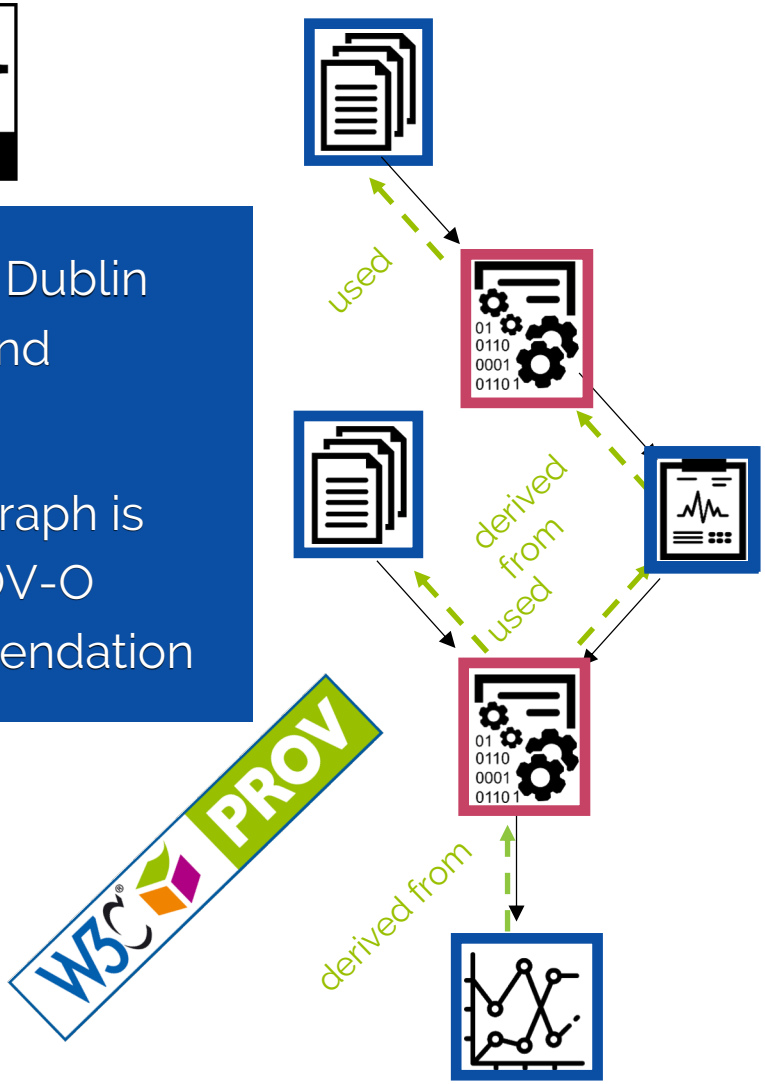


1. Inputs and outputs of analysis steps are recorded into a **knowledge graph** *as the work is being done*
2. Steps can be **repeated** or integrated into more complex **workflows**
3. **Provenance** of all data products is always accessible via simple tools
4. **Version control** is built-in for data,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

Encapsulate with rich metadata

JSON-LD

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

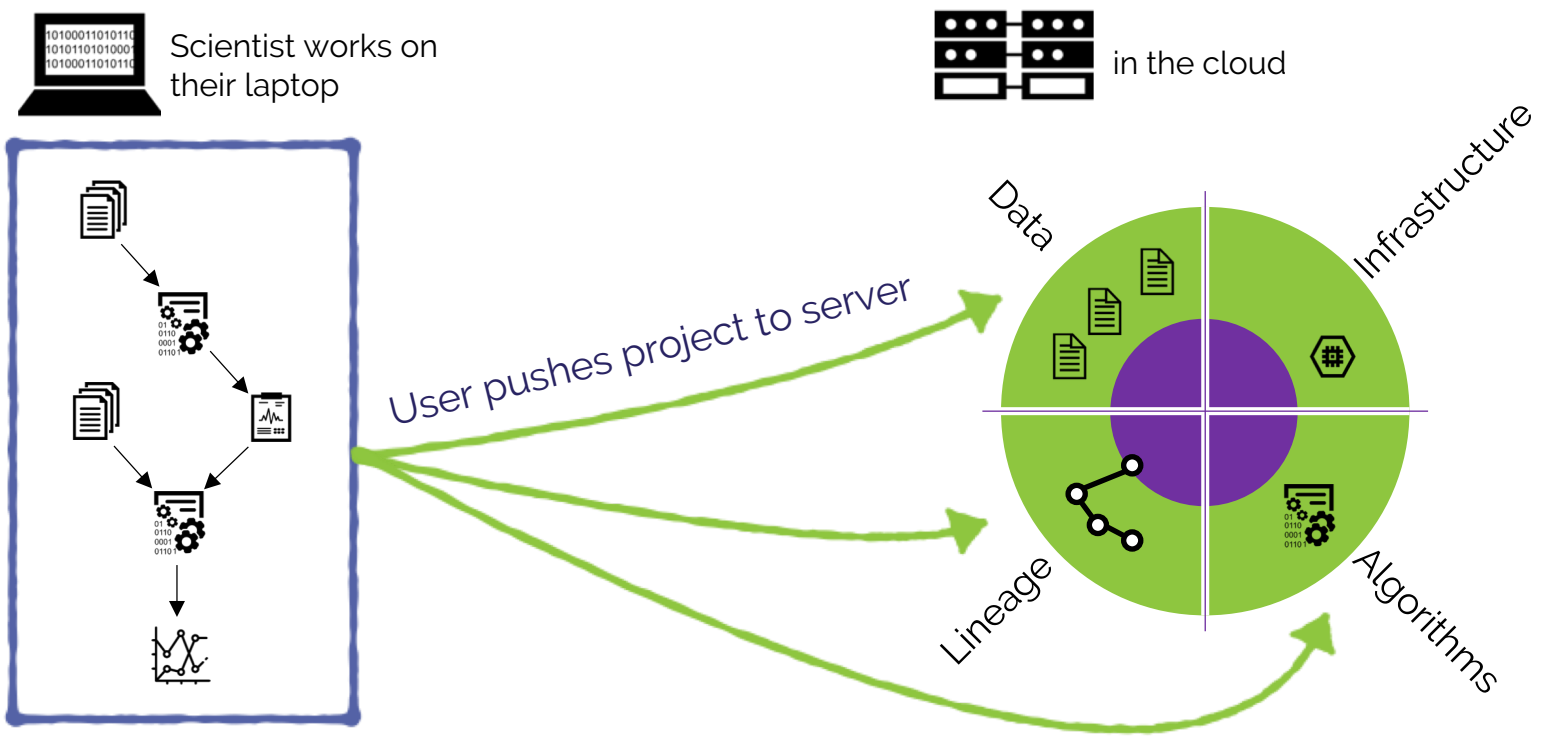


COMMON WORKFLOW LANGUAGE

- 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

SDSC

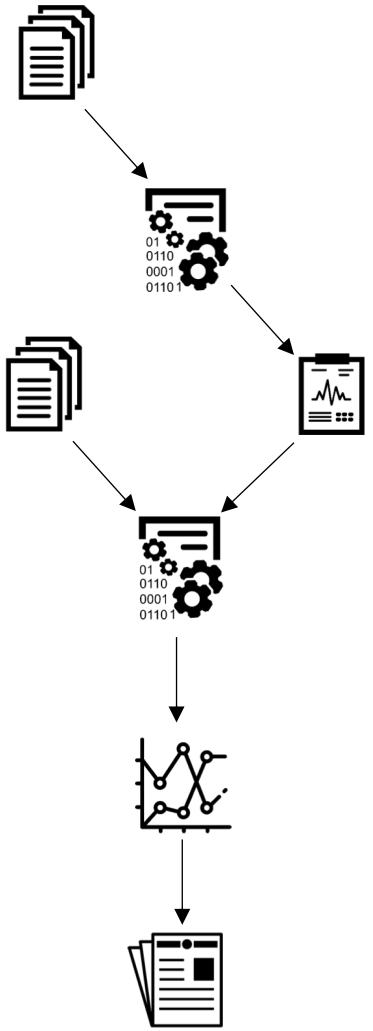
Verify and share results



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…

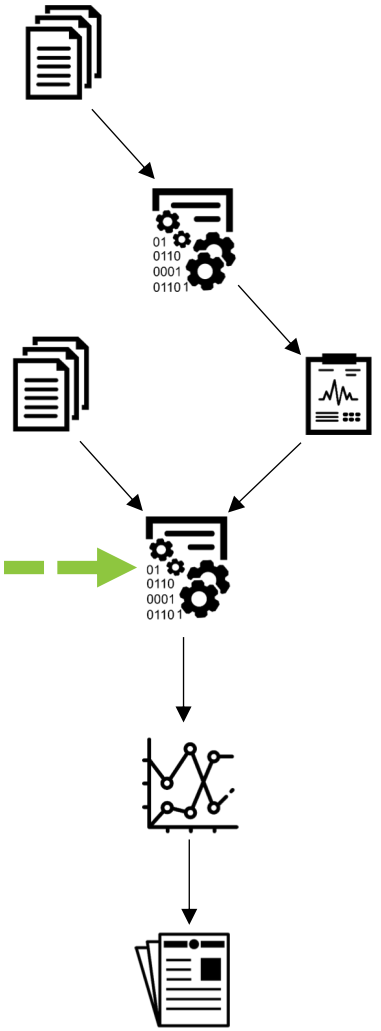


Ex: search for **data** and explore the tools to efficiently generate results

Discover and understand the work of others

Graph-based search…

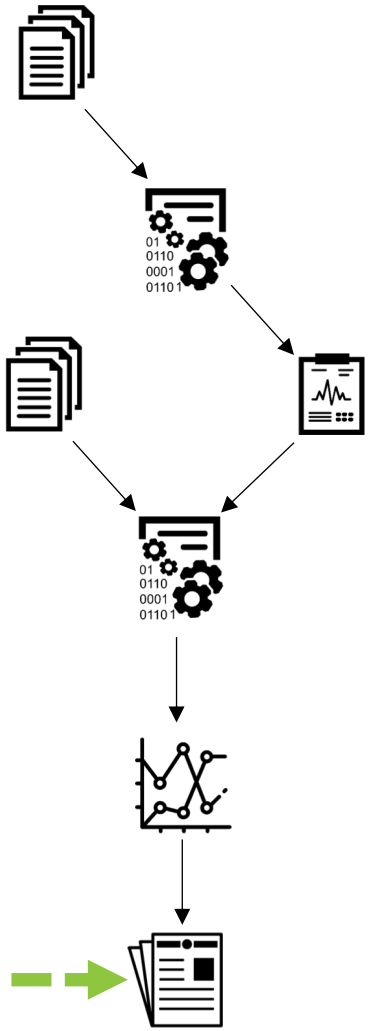
Ex: search for an **algorithm** and see its applications



Discover and understand the work of others

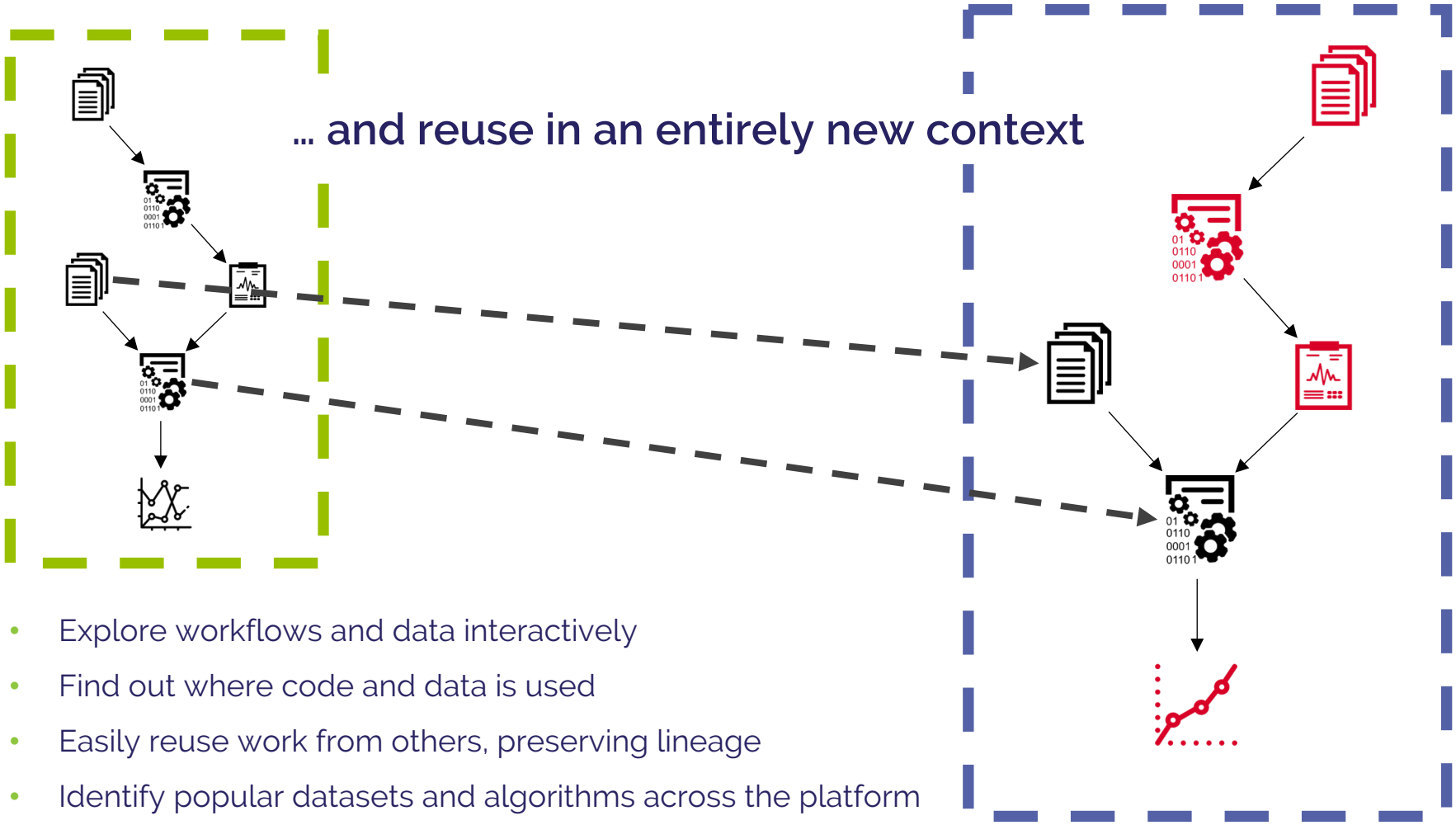
Graph-based search…

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



Reuse and repeat

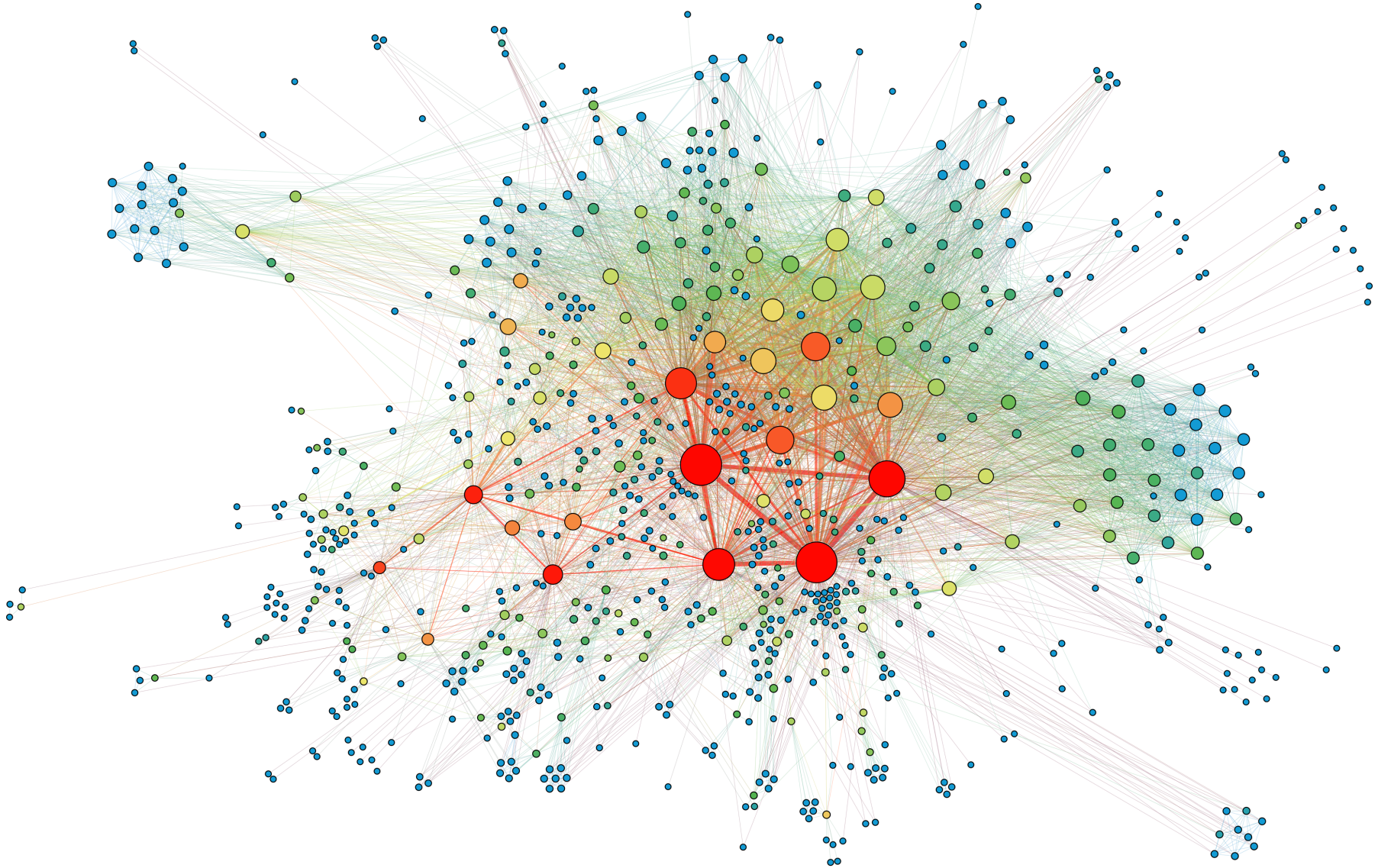
Search for relevant data or algorithms…



- Explore workflows and data interactively
- Find out where code and data is used
- Easily reuse work from others, preserving lineage
- Identify popular datasets and algorithms across the platform

SDSC

A Knowledge Graph emerges



SDSC

Vis. by Grandjean, M. 2015

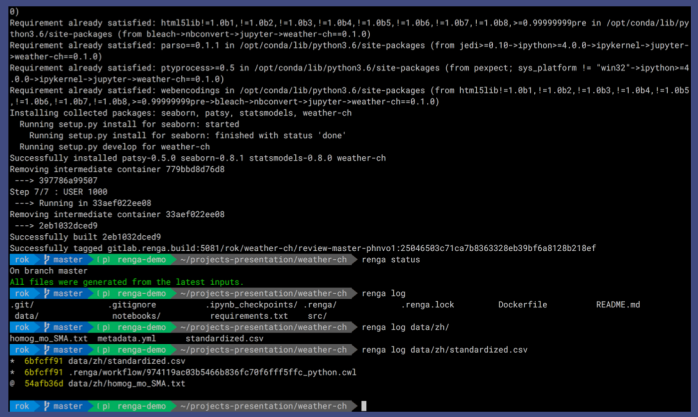
A peek into the data science process

- 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?

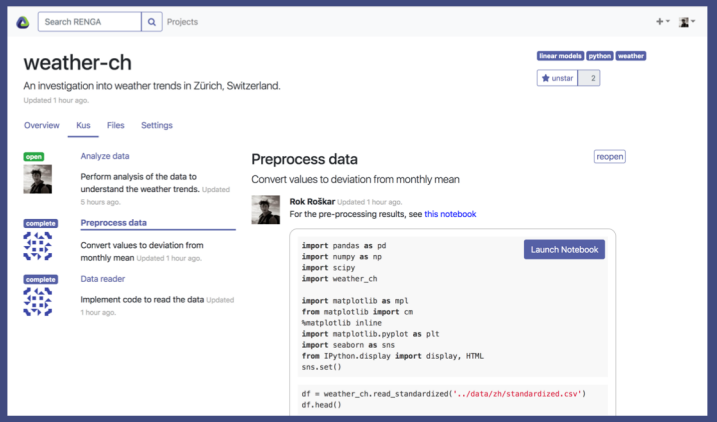
TRUST

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



Command-line interface



Web-based front-end

SDSC

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
```

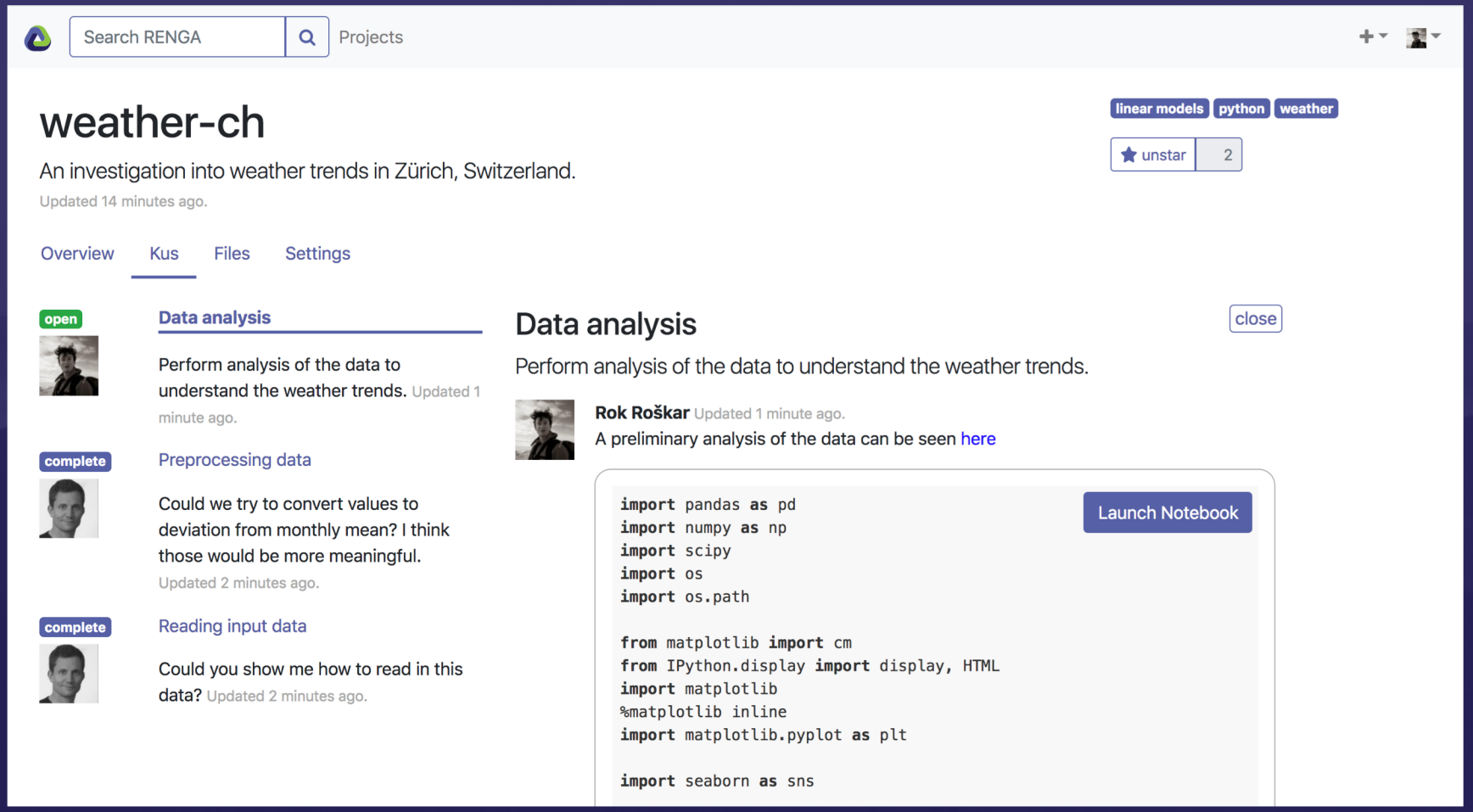
Lineage is captured and recorded in the local knowledge graph

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 <sha1> <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



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)