ETH-RAT

Enabling Reproducible, Repeatable and Reusable (Data) Science With the RENKU Platform

Rok Roškar | rok.roskar@sdsc.ethz.ch

SDSC

Swiss Data Science Center by the numbers

- ~30 people, scaling up to ~40 by mid-2019
  - ~10 on Renku platform/software engineering
  - ~15 data scientists from variety of backgrounds
- 18 academic collaboration projects
  - Funding mix of infrastructure and personnel
- several industry collaborations, scaling up with addition of a dedicated industry component

SDSC

Reproducibility — what do we mean?

- The “reproducibility crisis”*
  - Lack of replication studies
  - Failure to reproduce published results
  - Publication bias (only publishing “good” results)
  - False positives in published research
  - Lack of transparency and completeness

The upside: “credibility revolution”

* from the Stanford Encyclopedia of Philosophy

SDSC

Reproducibility — what do we mean?

- Reproduce studies based on the “concept”
- Repeat entire experiments following published methods/ protocols (but with own implementation)
- Repeat the published computational procedures on original or new data
- Repeat computations “verbatim” (same code, environment)

Five FAQs in Data-Driven Research

1. How did I compute this result?

2. How does new data change this result?

3. How did you compute *your* result?

Can I use your data to reproduce it?
With your code?
On your infrastructure?

4. Has anyone ever used an <XYZ-algorithm> on this data? How?

5. Who is using my data? and my algorithm?

Why are they not citing me?!

If you can answer these questions confidently:

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

Public focus

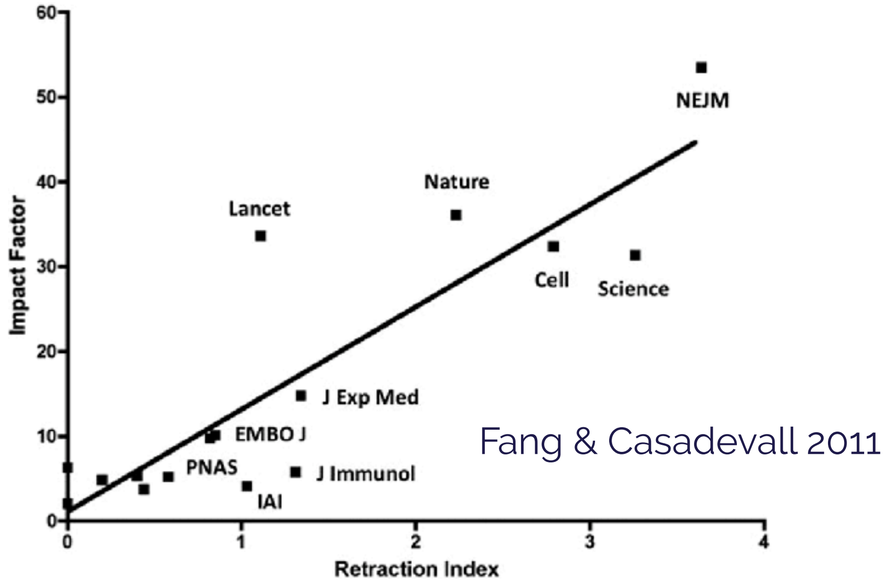
Kultur > Wissen >

Glaubwürdigkeit in Gefahr

Pimp my presentation: Wenn die Wissenschaft panscht

Viele Studien lassen sich nicht wiederholen. Das wirft ein schlechtes Licht auf die Wissenschaft. Was dagegen tun?

Montag, 03.09.2018, 15:33 Uhr

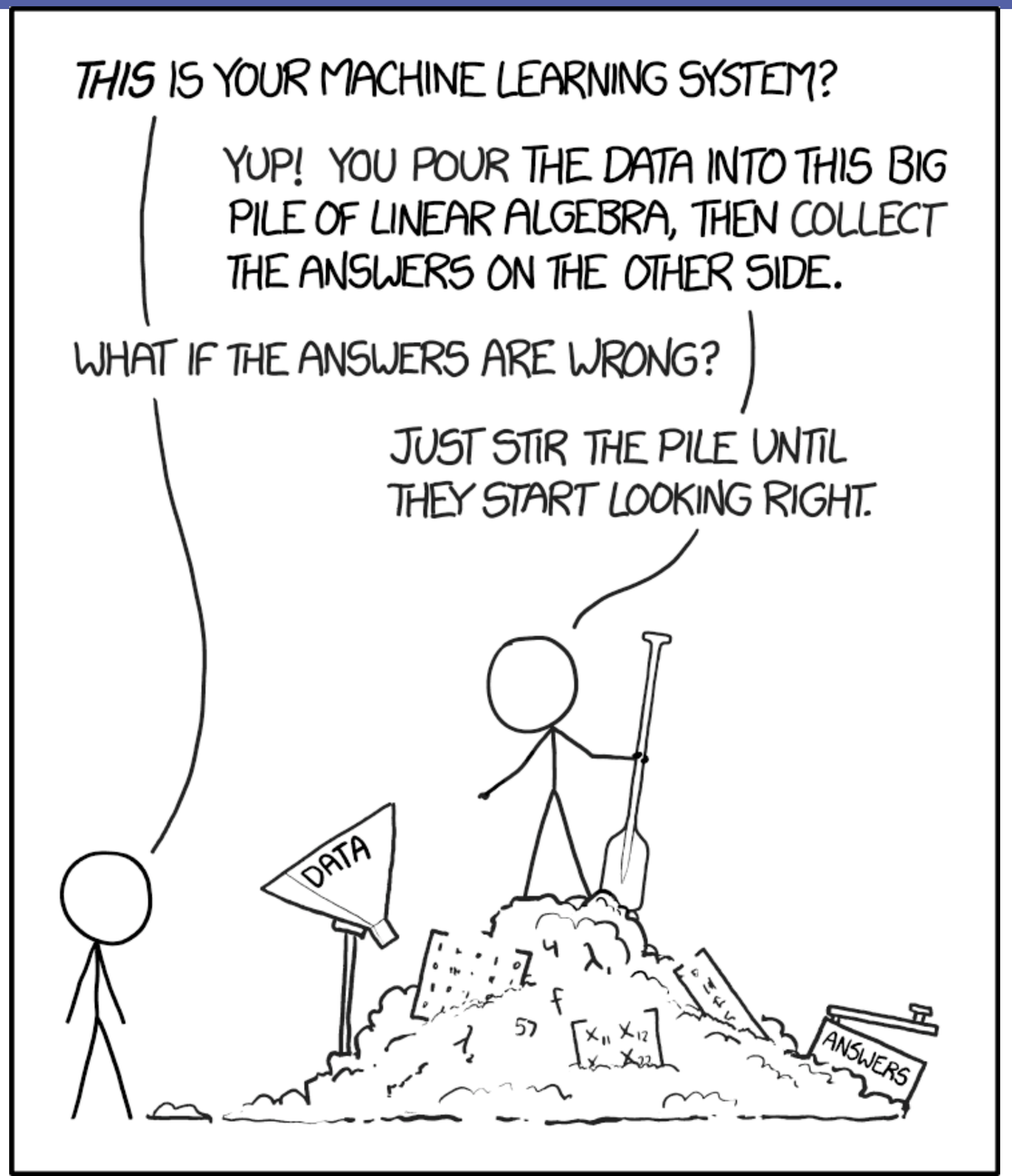


Fang & Casadevall 2011

- Increasingly, (failed) reproducibility in publicly-funded science under scrutiny
- Requirements for open science from funding agencies, governments (downward pressure)
- Some places (e.g. Switzerland) *openness* also across e.g. government data
  - But, is open useful in and of itself?
  - How can the open data be used?
  - How can the open science be reused? Reproduced?

SDSC

Situation to avoid...



Does “reproducible” mean “open”?

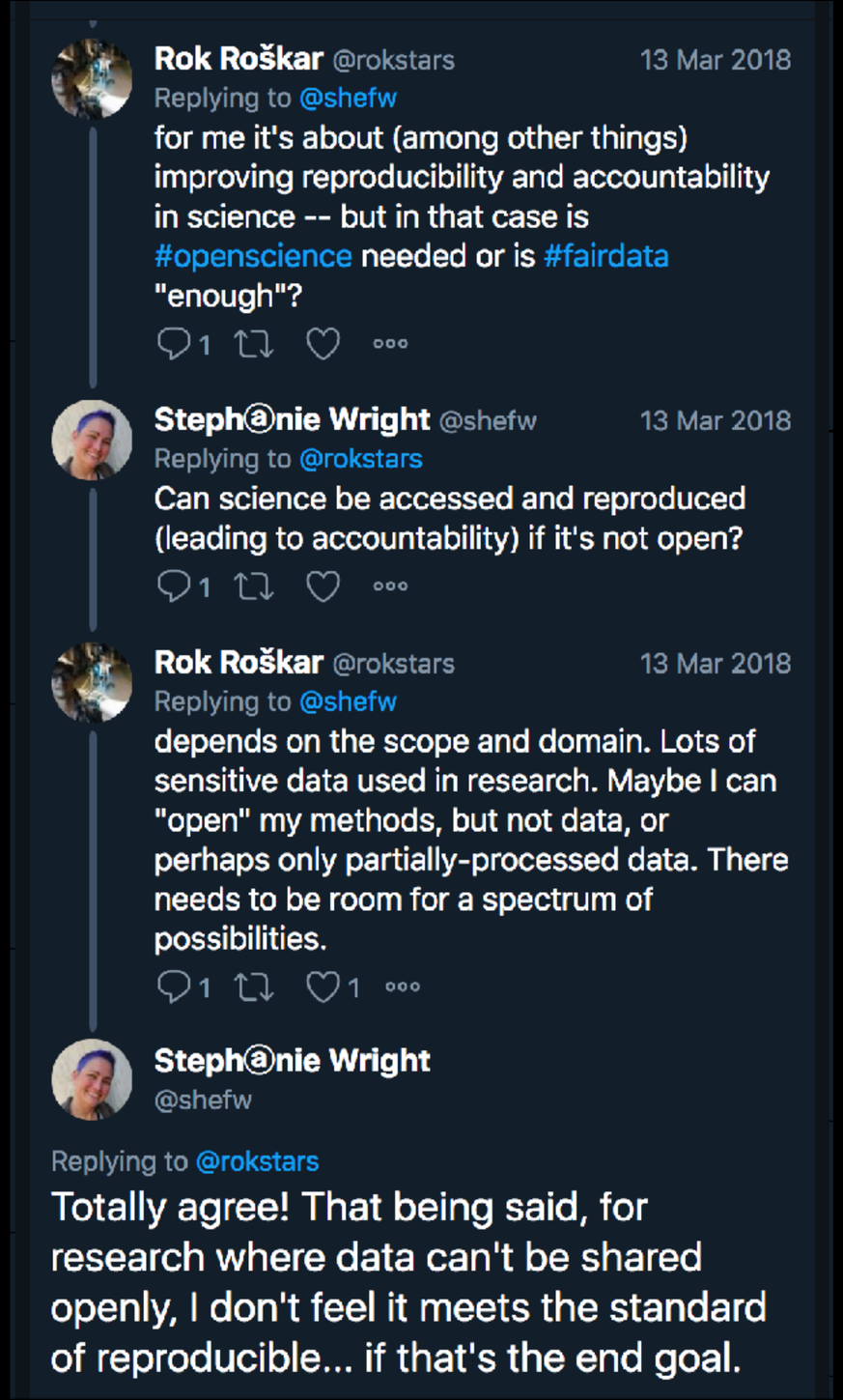
Does “open” mean “reproducible”?

Does FAIR mean “reproducible”?

Does FAIR mean “open”?

“reproducible” may depend on context and point of view…

Q: Which problem is Open Science fixing?



Rok Roškar @rokstars 13 Mar 2018
Replying to @shefw
for me it's about (among other things) improving reproducibility and accountability in science -- but in that case is #openscience needed or is #fairdata "enough"?
1

Steph@nie Wright @shefw 13 Mar 2018
Replying to @rokstars
Can science be accessed and reproduced (leading to accountability) if it's not open?
1

Rok Roškar @rokstars 13 Mar 2018
Replying to @shefw
depends on the scope and domain. Lots of sensitive data used in research. Maybe I can "open" my methods, but not data, or perhaps only partially-processed data. There needs to be room for a spectrum of possibilities.
1 1

Steph@nie Wright
@shefw

Replying to @rokstars
Totally agree! That being said, for research where data can't be shared openly, I don't feel it meets the standard of reproducible... if that's the end goal.

A Myriad of Tools

Hard **work** to make science reproducible, accessible, open etc…

- Version control (`git`)
- Code sharing (GitHub, GitLab, BitBucket)
- Data sharing (Zenodo, Figshare, institutional digital archives)
- Presentation and communication (Jupyter, RStudio)
- Correctness – testing (CI, e.g. travis)
- Packaging, containerization (docker, singularity)

Difficult to stay **productive** and worry about all of the above!

SDSC

FAIR + Open + ? Make it also practical!

- Findable, accessible, interoperable, reusable - data. What about the process?
- Top-down pressure for openness
  - It's in the public interest!
  - Is it in the interest of the individual researcher?
- Focus on the process also focuses on the scientist - making the process better will enable + incentivize
- **Scientists are practical - make "open" work for them!**

Reproducibility is critical here

SDSC

SDSC

RENKU - 連句

A Platform for Reproducible, Reusable, Shareable, (Data) Science

Terminology

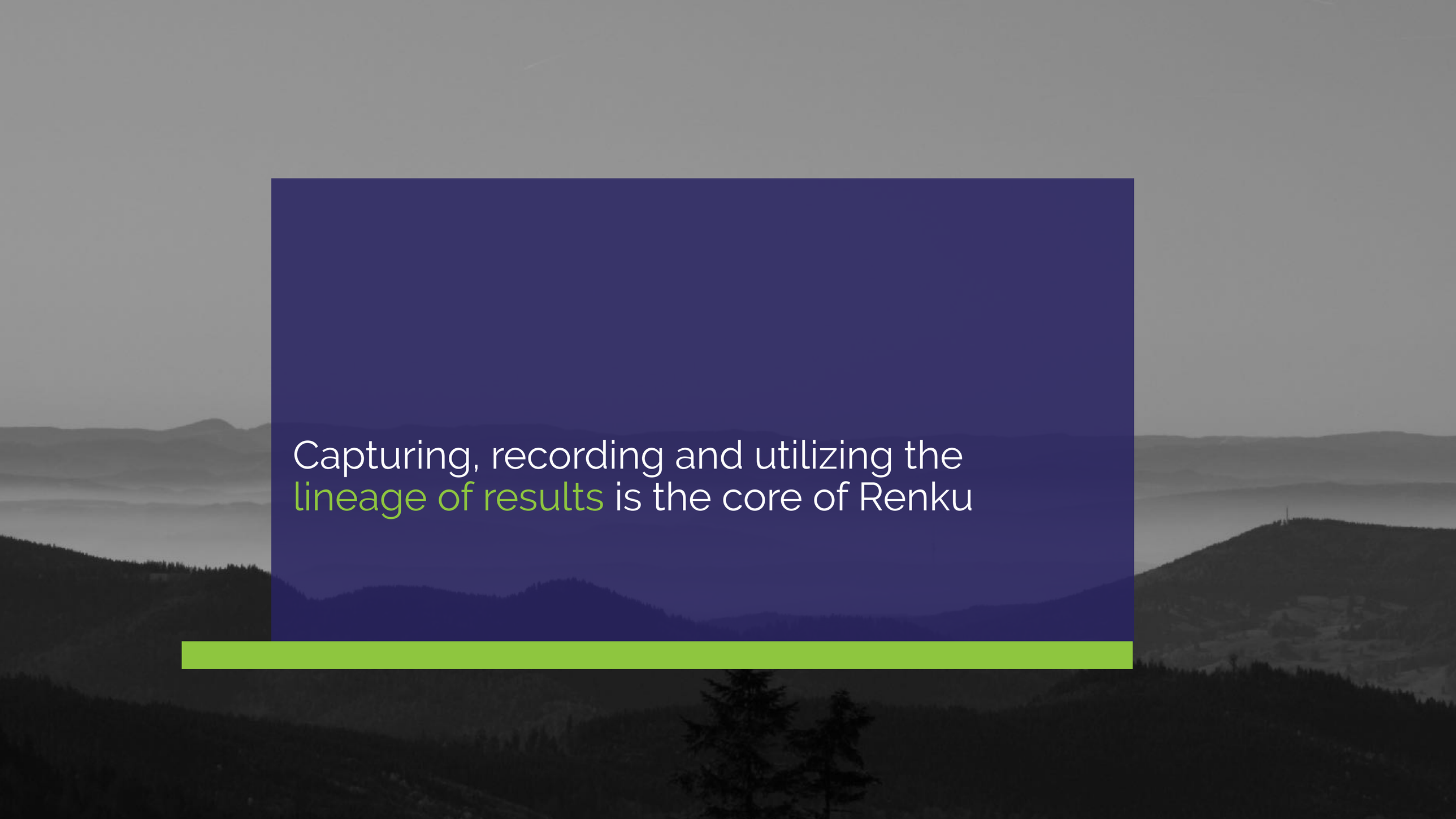
- We borrow the **Renku** name from the Japanese word for *linked-verse poetry*
- A "**ku**" is a verse in a renku poem
- We use "**ku**" to mean a piece of the data analysis process – includes discussion, code, and results

Five Questions → Three Words

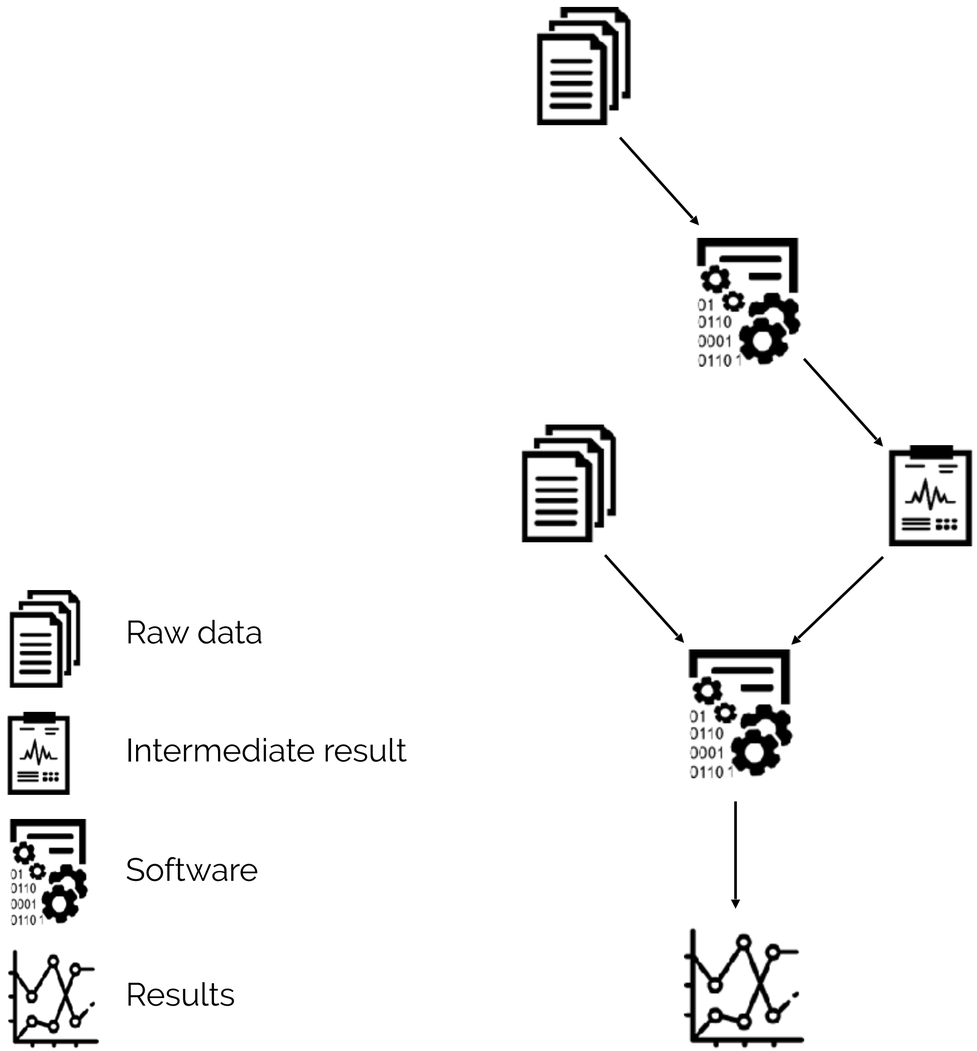
Reproducibility

Reusability

Collaboration


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

Lineage of a simple analysis

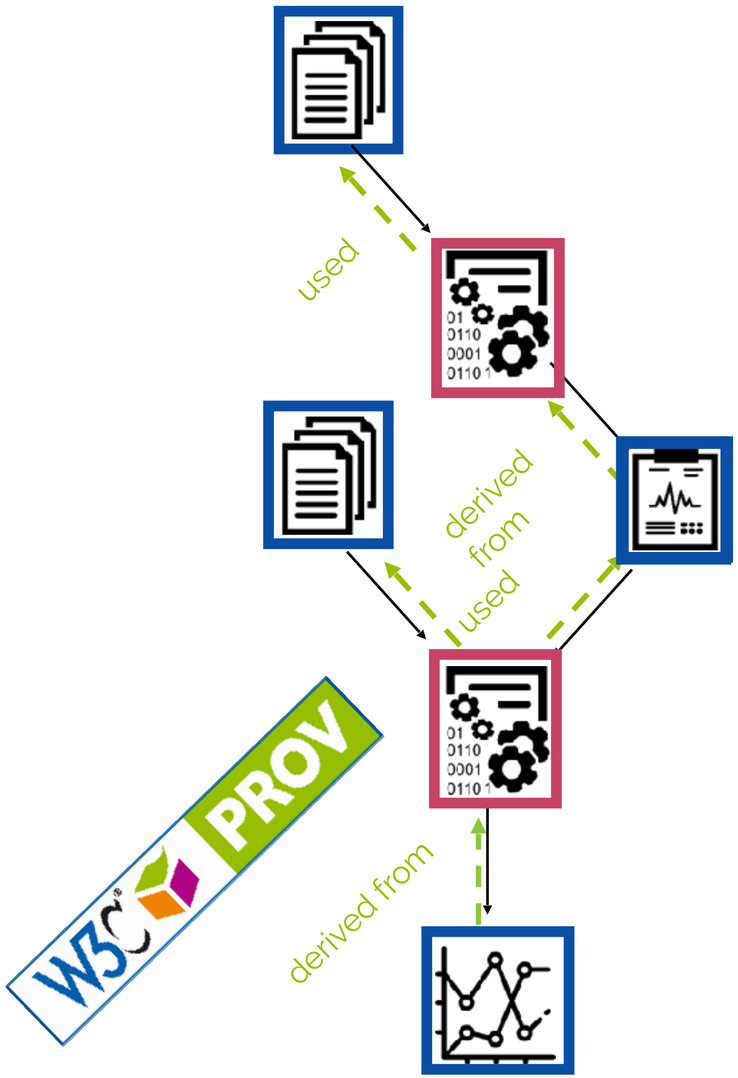


- Inputs and outputs of analysis steps are recorded into a **knowledge graph** *as the work is being done*
- Steps can be **repeated** or integrated into more complex **workflows**
- **Provenance** of all data products is always accessible via simple tools
- **Version control** is built-in for data,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

SDSC

Capturing the Knowledge Graph

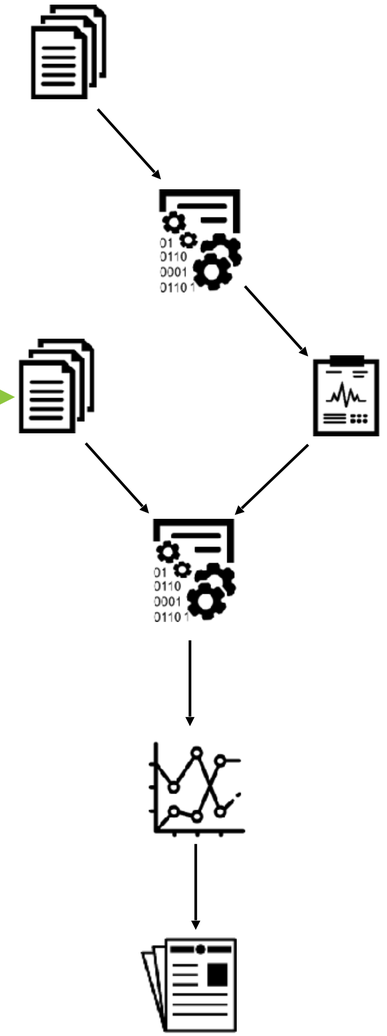
- Record the dependency graph using linked data standards e.g. Dublin Core, FOAF, Schema.org, PROV-O
- Capture the full semantic information to enable interoperability and domain-specific ontologies



- 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 consistent environment for reproducibility

Discover and understand the analysis process

Graph-based search…

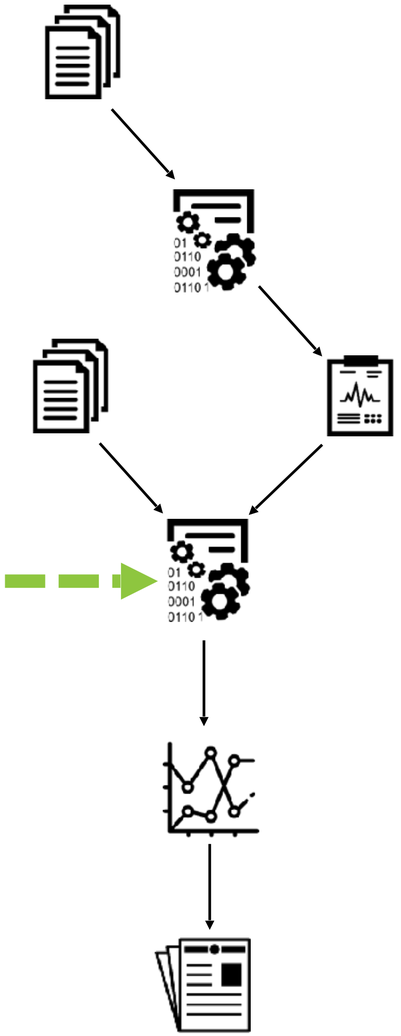


Ex: search for **data** and explore the tools to efficiently generate results

SDSC

Discover and understand the analysis process

Graph-based search…



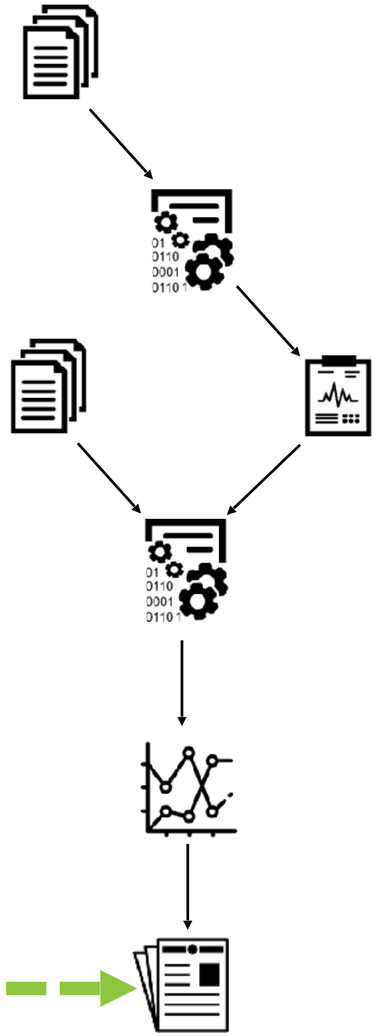
Ex: search for an **algorithm**
and see its applications

SDSC

Discover and understand the analysis process

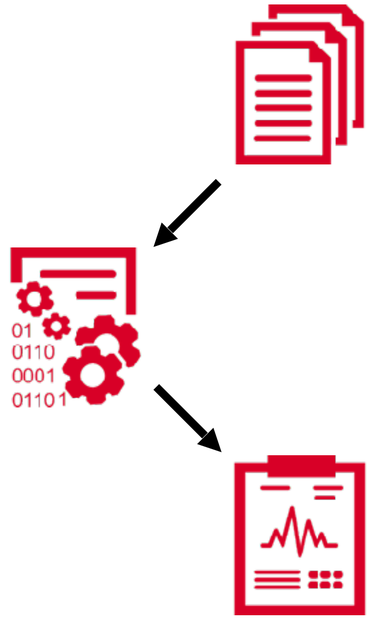
Graph-based search…

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



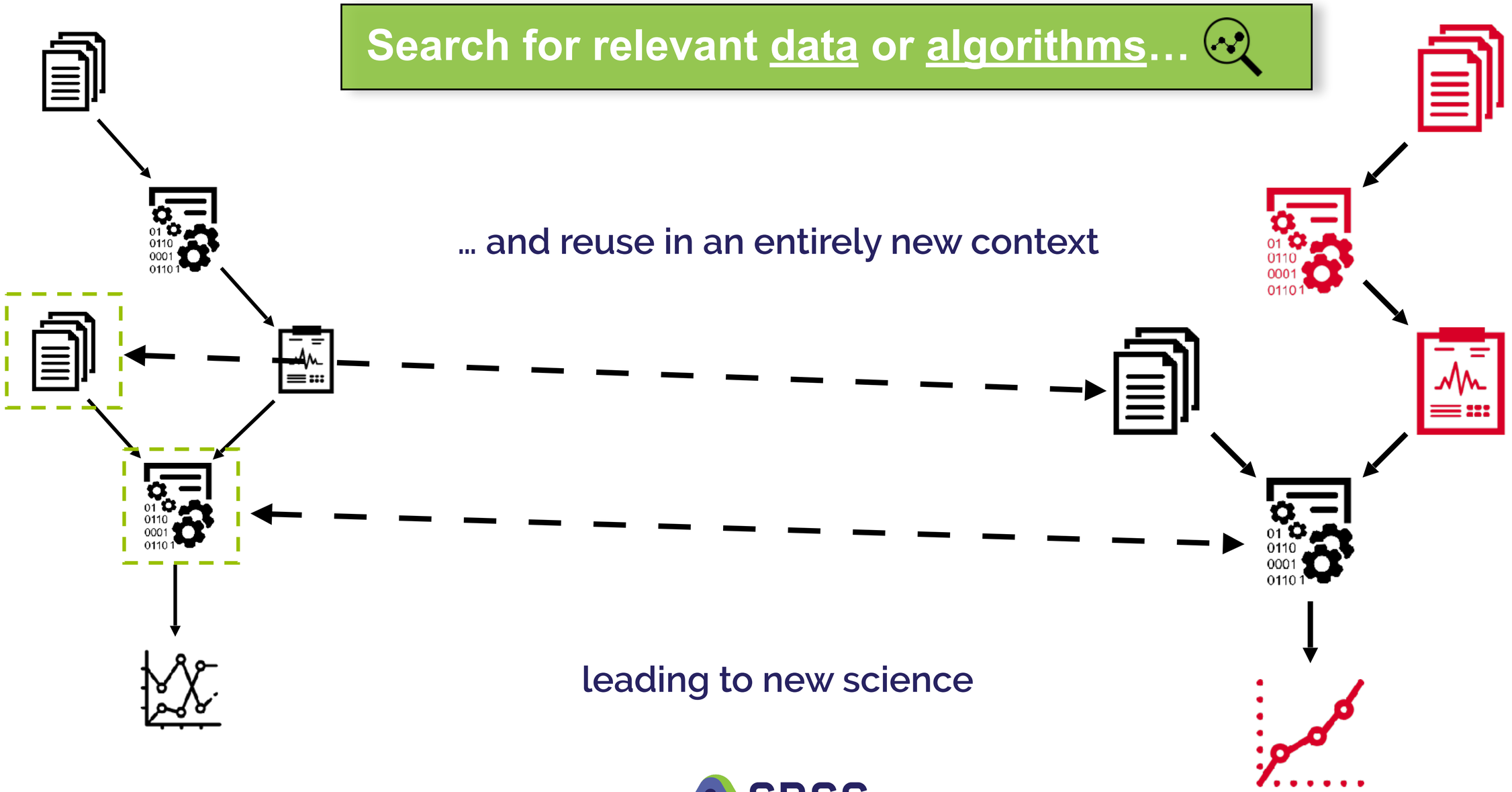
SDSC

Reuse and repeat



SDSC

Reuse and repeat



SDSC

Explore the KG to answer questions

- 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?

Best-practices translate to immediate benefits for the scientist

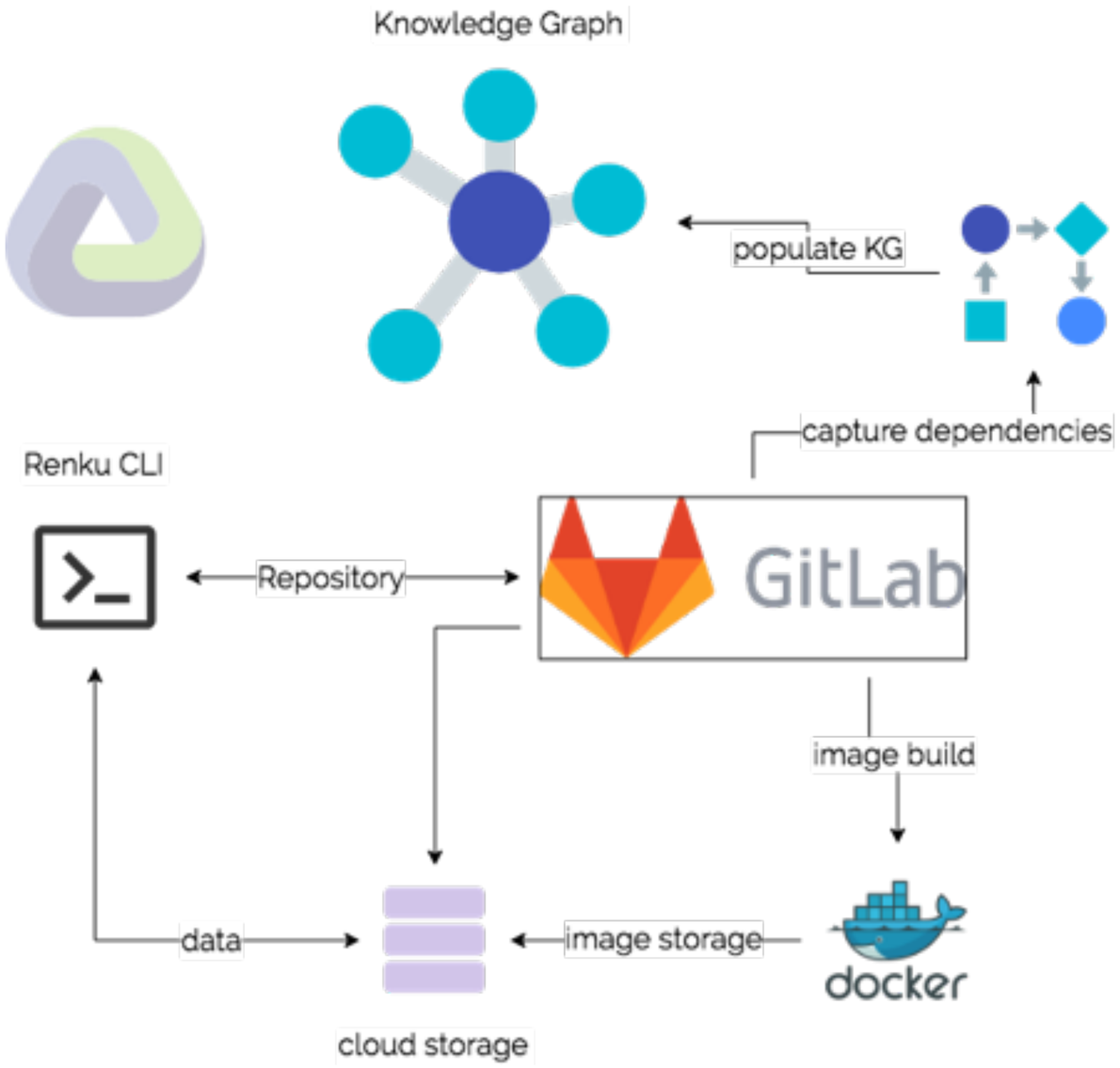
Establish trust

SDSC

Technology

- Core is based on trusted tech like git, GitLab, Jupyter, docker
- Easily deploys on any cloud
- Designed to be as independent/closed as needed, but as open as possible
- Example public deployment on renkulab.io - try the demo!

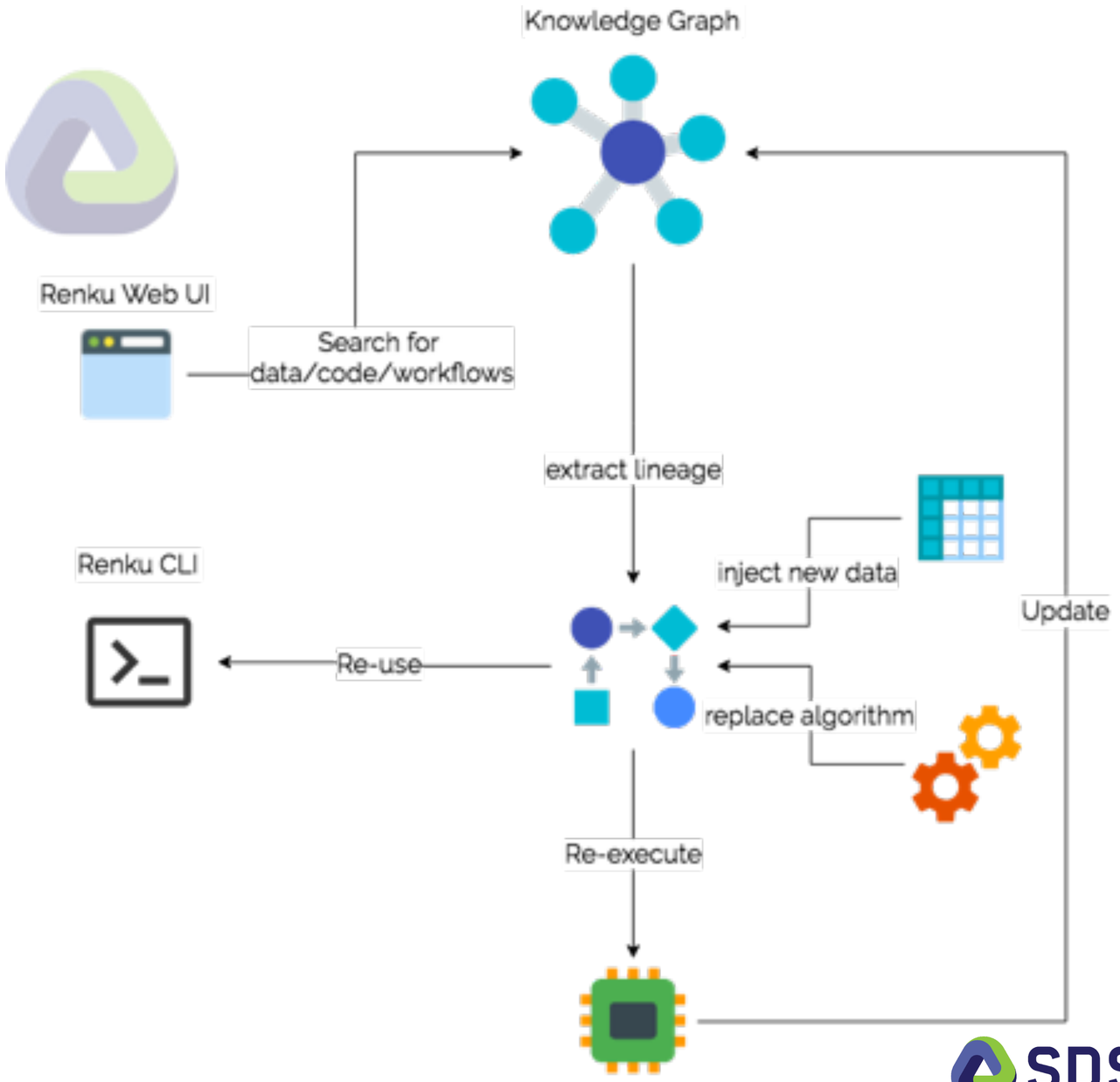
Reproducibility



- CLI developed from ground up taking advantage of Git to capture lineage
- Designed with interoperability in mind, using linked data standards to express dependencies
- Using version control for data and code – overcoming usability challenges

SDSC

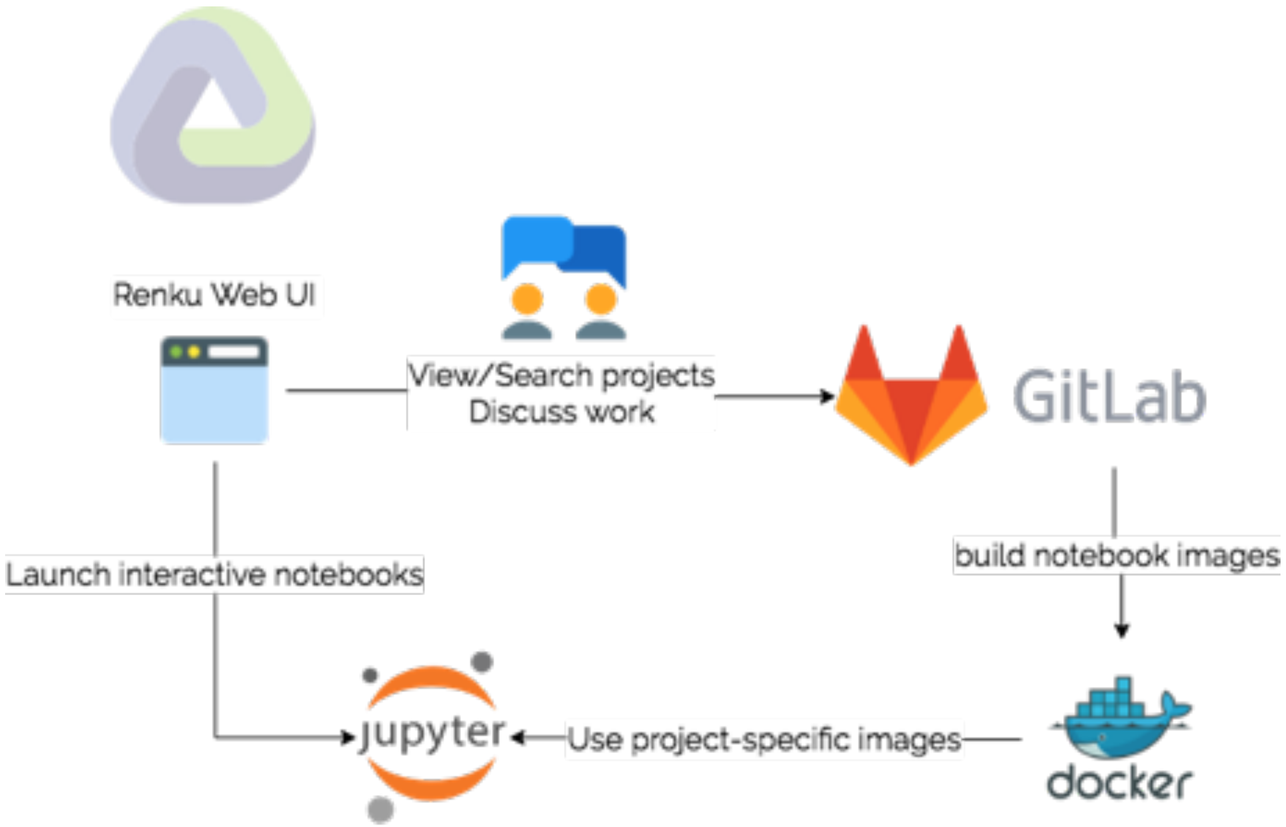
Reusability



- Understanding lineage means we can always re-execute and update results as methods or data change
- Workflow construction and re-execution possible with the local client
- Developing the means to execute workflows in the cloud and on HPC (CSCS)
- Search for graph artifacts being developed

SDSC

Collaboration



- Web UI serves as primary point of contact for users
- Allows creation of projects, discussions with media embedding
- Creation of hosted interactive sessions with version-controlled environments (docker images)
- Various supported languages, e.g. python, matlab (with GUI), R, Rstudio

SDSC

Renku platform

Goal: improve the scientific process by making use of best-practices simple

- Track the lineage of research artifacts
- Index lineage into a combined Knowledge Graph that can be queried
- Use the lineage through simple tools (e.g. update results with new data)
- Make all steps reusable and repeatable by others
- Provide a simple user interface for collaboration and exploration
- Enable on-line interactive environments for rapid prototyping

SDSC

Status and upcoming challenges

Status

- Platform is under very active development
- Releasing v0.4.0 within a few weeks - preview with tutorial at: https://renkulab.io
- All open-source: https://github.com/SwissDataScienceCenter

Challenges

- Knowledge representation/engineering
  - What questions will the users ask?
  - What is the optimal way to represent the graph (with FAIR+ in mind)?
- Community building and awareness
- Education - make scientists aware of and believe in “best practices”

SDSC